Universität Hildesheim
Fachbereich III: Informations- und Kommunikationswissenschaften
Institut für Angewandte Sprachwissenschaften
Studiengang Internationales Informationsmanagement (M.A.)

Magisterarbeit

Schwerpunkt Angewandte Informationswissenschaften

Informationslinguistische Ressourcen für das Information Retrieval in der tschechischen Sprache im Rahmen des Cross Language Evaluation Forums (CLEF)

Laura Hofman Miquel

Erstgutachterin:
Prof. Dr. Christa Womser-Hacker

Zweitgutachter:
Dr. Thomas Mandl

Hildesheim, Juli 2005

Abstract

Due to the effect of globalization and the increasing use of network-based systems, the situation of the search for information changed. The focus of interest switches to languages other than English. This work determines, analyzes and generates powerful information-linguistic resources for the information retrieval of the Czech language. As a result of this work a general stoplist for the Czech language and an intellectually built text-catalogue for the Czech toplevel-domain from WebCLEF are presented. Furthermore, this work includes the evaluation of the Polish stemmer STEMPEL. Its application for Czech texts is discussed.

Zusammenfassung

Durch die Globalisierung und den wachsenden Gebrauch von netzwerkbasierten Systemen hat sich die Situation für die Informationssuche geändert. Die englische Sprache verliert in diesem Kontext an Gewicht, sodass andere Sprachen in den Vordergrund rücken. In dieser Arbeit werden für die tschechische Sprache mächtige informationslinguistische Ressourcen bestimmt, analysiert und erstellt. Die Ergebnisse dieser Arbeit stellen eine allgemeine tschechische Stoppwortliste und einen intellektuell erstellten Text-Katalog für die tschechische Toplevel-Domain von WebCLEF dar. Weiterhin umfasst diese Arbeit die Evaluierung des polnischen Stemmers STEMPEL. Seine Anwendung für tschechische Texte wird kritisch betrachtet.

Inhaltsverzeichnis

Kapitel 1
Einleitung

Die rasante Entwicklung des Internets und elektronischer Datensammlungen, sowie die Fortschreitung der auf ihnen aufbauenden Informationssystemen hat in den vergangenen Jahren weltweit zu einem gewaltigen Anstieg der allgemein zugänglichen Wissensbestände geführt. Gleichzeitig führt diese Entwicklung aber auch eine Informationsflut mit sich, die zur alltäglichen Herausforderung für den heutigen Menschen geworden ist. Vannevar Bush, der schon sehr früh befürchtete, dass die wissenschaftlich relevanten Informationen explodierende Ausmaße annehmen würden, sagte bereits im Jahr 1945:

> "There is a new profession of trail blazers, those who find delight in the task of establishing a useful trail through the enormous mass of the common record."[1]

Im Sinne dieses Zitates befasst sich diese Arbeit mit den Möglichkeiten der Informationsfindung. Im Internet stehen dem Nutzer Suchmaschinen zur Verfügung. In Bibliotheken helfen ihm inzwischen meistens elektronische Katalogsysteme bei der Informationssuche. Die Suche in Dokumentkollektionen erfolgt durch *Information Retrieval-Systeme* (IR-Systeme). Diese ermöglichen den Zugriff auf Dokumente durch die Eingabe von natürlichsprachlichen Begriffen. Im Zuge der Globalisierung und Internationalisierung liegt das Wissen immer häufiger in verschiedenen Sprachen vor. Auch das Englische verliert nach und nach seine Rolle als „lingua franca“ des Internets. Bereits heutzutage sind 64,2% der Internutzer nicht-englischsprachig“.[2]

Die Informationssuche führt immer öfter über Sprachgrenzen hinweg, da sich die geeignete Information nicht immer im Wissensbestand des eigenen Sprachraums befindet. Der Bereich der IR-Forschung, der sich mit der Informationsbeschaffung über Sprachgrenzen hinaus befasst, ist das *Cross-Language Information Retrieval* (CLIR).

---

[1] *Bush, Vannevar (1945), As we may Think. The Atlantic Monthly.http://ccat.sas.upenn.edu/~jod/texts/vannevar.bush.html*

[2] http://global-reach.biz/globstats/index.php3

Um die Qualität von CLIR-Systemen kontinuierlich zu optimieren, wurde die europäische Evaluierungsinitiative *Cross-Language Evaluation Forum* (CLEF) gegründet. Bei CLEF sollen nicht nur die größeren Sprachen Europas getestet werden, sondern vermehrt auch die „kleinen Sprachen". Eine dieser Sprachen ist das Tschechische.

Gegenstand dieser Arbeit ist es, informationslinguistische Ressourcen für die tschechische Sprache mit dem Hintergrund eines zukünftigen Einsatzes in CLEF zu analysieren und zu erstellen.

Zu diesem Zweck wird in Kapitel 2 auf die Grundlagen des Information Retrievals eingegangen. Nach den Begriffsdefinitionen werden die gängigen Information Retrieval-Modelle und die Information Retrieval-Techniken erläutert. Dem folgen eine Schilderung der Evaluierungsmöglichkeiten für Information Retrieval-Systeme und ein Exkurs in das multilinguale Information Retrieval. Abschließend werden das MIMOR-Modell und das Cross-Language Evaluation Forum (CLEF) kurz vorgestellt.

Kapitel 3 geht zunächst auf allgemeine Aspekte der tschechischen Sprache ein, um darauf aufbauend die Besonderheiten der tschechischen Sprache, die im Information Retrieval-Prozess zu Problemen führen können, herauszuarbeiten.

In Kapitel 4 werden die bedeutendsten Forschungsinitiativen für informationslinguistische Ressourcen der tschechischen Sprache vorgestellt. Diese umfassen zum einen Aktivitäten an einzelnen Universit und zum anderen europäische Projekte.

Kapitel 5 widmet sich der Analyse und Erstellung von informationslinguistischen Ressourcen der tschechischen Sprache. Der Schwerpunkt der Analyse liegt bei der Bewertung dieser Ressourcen hinsichtlich der Integration in ein IR-System. Zunächst werden Korpora und die Arbeit mit ihnen vorgestellt. Ein Ergebnis dieser Arbeit sind die anschließend angeführten Stoppwortlisten. Danach folgt die Evaluierung des polnischen Stemmers STEMPEL, der im Rahmen dieser Arbeit auf tschechischen Input angewendet wurde. Abschließend folgt die Dokumentation der intellektuellen Erstellung eines Text-Kataloges für die tschechische Toplevel-Domain von WebCLEF.

Kapitel 2
Grundlagen

2.1 Information Retrieval

Information Retrieval (IR) ist ein interdisziplinärer Begriff, der je nach Sichtweise unterschiedlich interpretiert werden kann. Nach Womser-Hacker (2003) umfasst das IR im Rahmen der Informationswissenschaft

> „Verfahren, um Informationsobjekte zu erschließen und wiederaufzufinden".

Die Informationen sind also zunächst zu einem bestimmten Zeitpunkt nicht zugänglich und müssen durch das *Retrieval* (dt. „wiedererlangen") wieder aufgefunden werden. Die eben genannte allgemeine Definition kann mit der Definition von Salton und McGill (1987, 1) ergänzt werden, die besagt, dass im IR

> „die Repräsentation, Speicherung und Organisation von Informationen und der Zugriff zu Informationen"

zusammenfallen. Gemäß dieser Definition zählen zum IR auch das System und die Methoden des Bibliothekswesens, die mit Hilfe des Karteikartensystems die Bücher formal (durch beispielsweise Autor, Titel oder ISBN) und inhaltlich (durch beispielsweise Schlagwörter oder Abstracts) beschreiben. Die Suche in den Bibliotheken hat sich in den letzten Jahren stark verändert. Die Karteikartensysteme werden digitalisiert und durch computerbasierte Suchsysteme abgelöst, sodass der Informationssuchende über das Internet auf viele verschiedene Bibliothekskataloge zugreifen kann. Einige Bibliotheken existieren sogar nur als *digitale Bibliotheken*, das heißt, dass kein realer Bestand vorhanden ist.

Die Definition von Meadows (1991, S.2) verdeutlicht die Tatsache, dass IR nicht zwangsläufig an die Maschine gebunden ist und hebt den Kommunikationsprozess hervor:

> „IR is a communication process. [...] authors or creators of records communicate with readers, but indirectly and with a possibly long time lag between the creation of a message or text and its delivery to the IR system user. [...] Is information retrieval a computer activity? It is not strictly necessary that it be [...]".

Eine technischere Sichtweise vertritt Mayfield (2002). Er beschreibt den Kern des IR als

> „the automatic identification of those documents in a large document collection that are relevant to an explicitly-stated information need".

In dieser Definition wird, im Unterschied zu den bereits genannten Definitionen, die *automatische Identifizierung* in den Vordergrund gehoben und die Begriffe *Relevanz* und *explizit formuliertes Informationsbedürfnis* (Suchanfrage) werden eingeführt. Dieser computerunterstützte Bereich des IR nimmt stetig zu, da immer mehr Dokumente in digitaler Form vorliegen, wie z.B. im *WWW*. Aus diesem Grund wird der Begriff IR hauptsächlich mit der computerunterstützten inhaltsorientierten Suche assoziiert.

Weiterhin spricht Mayfield hier von *Dokumenten* (bisher wurden die Begriffe *Information* oder *Informationsobjekte* stattdessen verwendet). Um eventuelle Missverständnisse zu vermeiden, folgt nun zunächst eine Klärung des Begriffs *Information*. Der informationswissenschaftliche Informationsbegriff wird von dem Begriff *Wissen* differenziert. Wissen ist dabei

> „der Bestand an Modellen über Objekte und Sachverhalte der Welt, die in Individuen, in Gruppen etc. vorhanden sind" (Womser-Hacker 2003)

und wird im Problemlösungsprozess in Information umgewandelt. Den Zusammenhang zwischen den beiden Begriffen veranschaulicht die Formel: „Information ist Wissen in Aktion“ (Womser-Hacker 2003).

Zusammenfassend kann man sagen, dass Information auf Wissen basiert, sie aber an die Situation und den Kontext des Benutzers angepasst werden muss. Diese „Anpassung“, bzw. „Formgebung“ der Information ist bereits in der ursprünglichen Bedeutung des Wortes *informieren* zu finden. Das lateinische *informare* bedeutet so viel wie „eine Gestalt geben, formen, bilden“ (DUDEN, Das Herkunftswörterbuch 1997, 305). Information kommt in einem Kommunikationsprozess zustande, der an einen spezifischen Problemlösungszusammenhang gebunden ist. Sie ist also zielgerichtet. Des Weiteren wird Information nach ihrem Neuigkeitswert und ihrer Handlungsrelevanz bewertet und muss mengenmäßig angepasst werden (Informationsflut vs. Informationsdefizit). Es gilt die Daumenregel: *Soviel wie nötig, so wenig wie möglich* (vgl. Krcmar 2000, 6 ff.).

Die Aussage von Jürgen Rüttgers, ehemaliger BMBF, (In: Womser-Hacker 2003) „Information ist der Rohstoff für Innovation“ begründet die Motivation für das IR mit Informationen zu arbeiten.

Die folgende Definition von der (Fachgruppe IR 1996) beschreibt die zwei Grundprobleme *Vagheit* und *Unsicherheit*, die das Konzept des IR kennzeichnen und von der Suche in herkömmlichen Datenbanken abgrenzen:

> „IR beschäftigt sich schwerpunktmäßig mit jenen Fragestellungen, die im Zusammenhang mit vagen Anfragen und unsicherem Wissen entstehen. Vage Anfragen sind dadurch gekennzeichnet, dass die Antwort nicht a priori *eindeutig* definiert ist.“

Die Vagheit bezieht sich auf das schwer abzugrenzende und folglich nicht präzise und formal ausdrückbare Informationsbedürfnis des Benutzers. Die gesuchte Information ist dem Benutzer schon zu einem bestimmten Grad bekannt. (In relationalen Datenbanken hingegen gäbe es die Möglichkeit, beispielsweise durch SQL eine formale und somit für das System eindeutige Anfrage zu formulieren).

Im Gegensatz zur Vagheit bezieht sich das Problem der Unsicherheit auf das Wissen, das heißt die Kenntnisse des Systems über den Inhalt der Dokumente. Der Inhalt kann dabei verschiedene Formen annehmen. Zum größten Teil sind es Texte wie z.B. Zeitungsartikel und Berichte. Er kann aber auch z.B. Fakten, chemische Strukturen, Bild-, Audio- oder Videodokumente umfassen. Die gesuchte Information befindet sich dabei in den Dokumenten. Daher finden die meisten IR-Systeme, trotz ihrer Bezeichnung, im Grunde keine Informationen, sondern Dokumente.

Das Ziel des IR-Systems (IRS) ist es, die gespeicherten und später vom Benutzer gesuchten Informationen so aufzubereiten und in einer für den Benutzer angemessenen Form anzubieten, dass sie

> "bei einem konkreten Informationsbedarf mit problemangepaßten Suchstrategien und -operatoren interaktiv möglichst präzise (...) und vollständig herausgesucht werden können“ (vgl. Knorz 1995, 244).

Die grundsätzliche Aufgabe eines IRS basiert auf dem Vergleich von Benutzeranfragen mit den in einer Datenbank gespeicherten Dokumenten. Genauer gesagt werden die jeweiligen Repräsentationen verglichen. Dieser Prozess wird auch als *Match (*oder auch *Matching)* bezeichnet und wird in der Abb. 1 dargestellt.

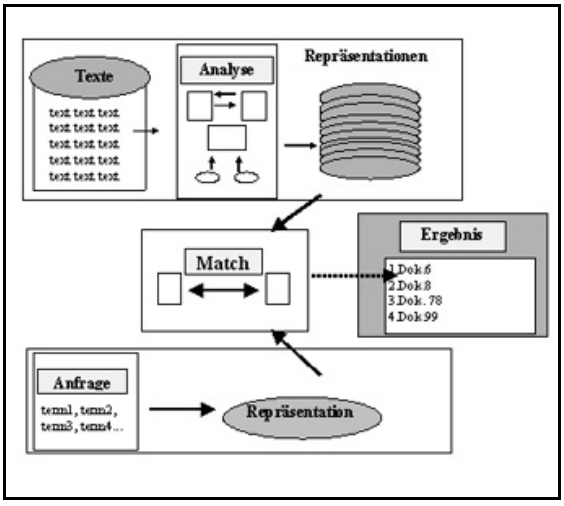

Abb. 1: *Grundmodell IRS (Womser-Hacker 2003)*

Der Vergleich gestaltet sich folgendermaßen: Auf der einen Seite werden die Dokumente analysiert, bearbeitet und anschließend in eine bestimmte Repräsentationsform gebracht und gespeichert. Auf der anderen Seite wird durch den Nutzer eine Suchanfrage mit einer bestimmten Anzahl von Termen gestellt. Die Anfrage wird ebenfalls in eine bestimmte Repräsentationsform gebracht. Das IRS vergleicht nun die Repräsentation der Anfrage mit derjenigen der Texte. Wenn diese übereinstimmen, liefert das IRS die zugehörigen Dokumente als Ergebnis. Dieses Ergebnis präsentiert sich dem Nutzer in den meisten Fällen als eine geordnete Liste (eine sog. *gerankte Liste*), in der die Dokumente nach *Relevanz* sortiert sind. Auf diese wird im Kapitel 2.5.3 genauer eingegangen.

In allen klassischen IR-Systemen werden die Dokumente intern vereinfacht bzw. normalisiert repräsentiert. Die Verallgemeinerung beinhaltet im Wesentlichen die Reduzierung und Normalisierung der Inhalte sowie die Indexierung und Kategorisierung, basierend auf unterschiedlichen Verfahren. Diese Verfahren werden im Kapitel 2.3 erläutert.

Folgende Graphik integriert den *Matching-Prozess* in den Gesamtzusammenhang des IR:

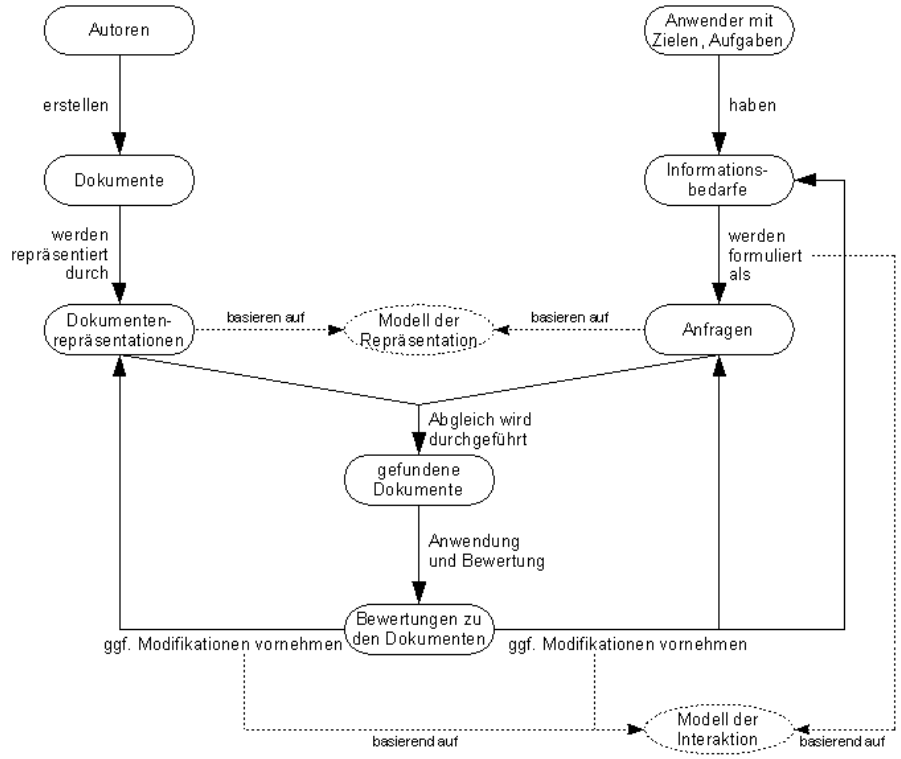

Abb. 2: *Ein allgemeines Modell zum Information Retrieval (Kuropka 2004, 9)*

Der IR-Prozess ist ein zyklischer Prozess, da während der einzelnen Phasen zu einer früheren Phase zurückgesprungen werden kann (z.B. aufgrund neuer Erkenntnisse bzgl. der Problemanalyse oder der Suchanfrageformulierung). Diese Eigenschaft ist ebenfalls der Abb. 2 zu entnehmen.

2.2 Information Retrieval-Modelle

Die Klassifikation von IR-Modellen und deren Detailliertheitsgrad variiert je nach Literatur und Perspektive. Das ausführliche und vielschichtige Modell von Belkin und Croft (1987, 112) hat mit dem folgenden Modell (Abb. 3) gemeinsam, dass die klassischen IR-Systeme in zwei Kategorien von Modellen eingeteilt werden. Die erste Kategorie wird von Modellen mit exakter Übereinstimmung gebildet (auch *Exact-Matching* genannt), die zweite von Modellen mit bestmöglicher Übereinstimmung (*Partial-Matching*), die sich wiederum noch genauer unterteilen lassen. Gemeinsam haben alle Verfahren, dass die Daten immer im Hinblick auf eine spätere Suche modelliert werden.

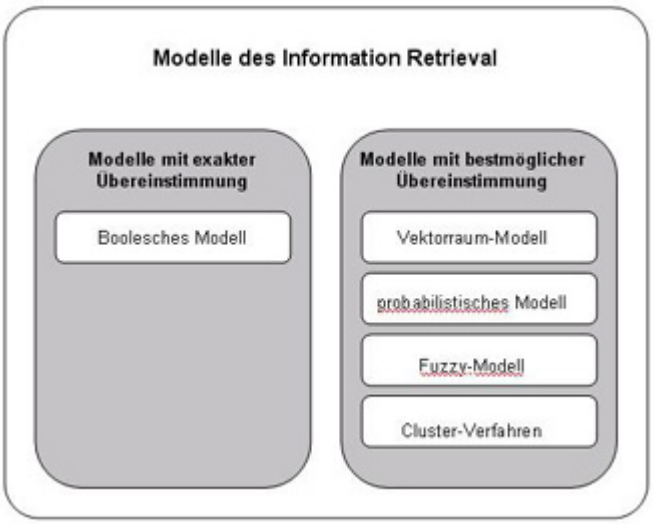

Abb.3: *Klassifikation der IR-Modelle*

Jedes IR System basiert auf einem oder mehreren dieser Grundkonzepte, die sich nicht unbedingt gegenseitig ausschließen.

2.2.1 Exact-Matching-Modelle

IR-Systeme, die auf einem Modell mit exakter Übereinstimmung basieren, ermitteln für jedes Objekt einen Wahrheitswert (wahr = 1, falsch = 0), der angibt, ob der exakte Term im Dokument vorkommt oder nicht und das Dokument somit für eine Anfrage relevant ist. Es finden keine Ähnlichkeitsbetrachtungen statt, da ein Dokument eine Suchfrage entweder befriedigt oder nicht. Modelle mit exakter Übereinstimmung sind mathematisch einfach und schnell in der Ausführung.

Eines der bekanntesten Modelle mit exakter Übereinstimmung ist das *Boolesche Modell*. Es ist das am einfachsten zu handhabende Retrieval-Modell. Im klassischen Booleschen-Retrieval-Modell drücken Benutzer ihr Suchproblem in einer exakten Retrievalsprache aus, wobei die Terme mittels der Operatoren der Booleschen Logik zu einer logischen Funktion verbunden werden.

Die Erstellung der Anfrageformulierung ist jedoch umständlich und kann den Benutzer überfordern, da er die Anfrage in der formalisierten Form mit Hilfe von *Booleschen Ausdrücken* eingeben muss. Die *Booleschen Ausdrücke* (Kombination von Termen, die über Boolesche Operatoren miteinander verknüpft sind) entsprechen nicht völlig der Semantik der natürlichen Sprache. So bedeutet bspw. das natürlichsprachliche „oder" meistens „entweder oder". Es schließt also aus, dass beide Möglichkeiten gelten, im Unterschied zu dem Booleschen Operator ODER, der als Ergebnismenge beide Mengen liefert.

Das Boolesche Modelle weist weiterhin folgende Nachteile auf:

- Die *disjunkte* (unvereinbare) Unterteilung der Dokumente in relevant und nicht-relevant lässt keine Rangordnung zu. Das Ergebnis kann folglich nicht nach Relevanzwerten sortiert werden.
- Beim Booleschen Retrieval existiert keine Möglichkeit für die Gewichtung der Anfrage, d. h. die einzelnen Suchbegriffe haben die gleiche Wertigkeit. Auch die Indexierung wird nicht gewichtet, sodass die einzelnen Indexierungsterme ebenfalls die gleiche Wertigkeit haben. In der Praxis tritt allerdings oft der Fall ein, dass sich ein Dokument hauptsächlich mit einem Beschreibungsterm beschäftigt und die anderen Beschreibungsterme nur eine untergeordnete Rolle

spielen. Daraus ergibt sich die Konsequenz, dass das Boolesche Modell nicht dem Bedarf des Nutzers entspricht.

- Der erwünschte Umfang der Ergebnisse ist schwer kontrollierbar und die Ergebnisse sind aufgrund der fehlenden Relevanzwerte schwer zu visualisieren.
- Der entscheidende Nachteil des Booleschen Retrievals ist sein Prinzip des "exact match" - dabei berücksichtigt es in keiner Weise die Vagheit der Anfrage (Der Benutzer kann nicht wissen, welche Indexierungsterme verwendet wurden.) bzw. die Unsicherheit der Repräsentation (Es ist unklar, ob alle notwendigen Begriffe zur Dokumentrepräsentation herangezogen wurden).

2.2.2 Partial-Matching-Modelle

Bei Partial-Matching-Modellen werden den Objekten der Datensammlung bei einer Suchanfrage keine Wahrheitswerte zugewiesen, sondern Werte, die einen Vergleich der Objekte anhand ihrer Relevanz zulassen. Das Ergebnis dieses Vergleichs ist eine Relevanzrangfolge, eine sog. *Ranking-Liste*. Die Vorteile von Ranking-Verfahren sind, dass durch die Rangordnung die relevantesten Dokumente an den Anfang der Liste gereiht werden und dass der Benutzer selbst das *cut-off* (Abbruch*)* bestimmen kann. So können Probleme mit großen Datenmengen vermieden werden. Eine für das Ranking-System notwendige Voraussetzung ist eine gewichtete Indexierung der Dokumente. Bereits für sehr einfache Verfahren zeigen Experimente bei der Verwendung von Ranking-Verfahren eine bessere Retrievalqualität (vgl. Salton, McGill 1987, 156).

Im Folgenden sollen die wichtigsten Modelle, die nach dem Partial-Matching-Verfahren vorgehen, vorgestellt werden.

Bei der *Vektorraumsuche* werden Anfragen und Dokumente als mehrdimensionale Vektoren im Raum betrachtet. Dieser weist ebenso viele Dimensionen wie Indexterme auf (vgl Abb. 3). Das *Matching* entspricht einer Distanzbestimmung zwischen den Vektoren (vgl. Frakes, Baeza-Yates 1992, 366). Im *Vektorraum-Modell (VRM)* können Dokumente auf eine Anfrage teilweise relevant sein. Die Ergebnismengen werden absteigend nach Relevanz sortiert.

Zunächst werden zur Berechnung der Ähnlichkeit (oder Nähe) der Dokumente mit der Suchanfrage Vektorpaare gebildet. In diesem Schritt wird ermittelt, ob ein Term in

einem Dokument vorhanden ist. Bei der einfachen Anfrage mit den Termen t_1 und t_2 und den Dokumenten d_1 mit den Indextermen t_1, t_3, t_4 und d_2 mit t_1, t_2, t_4 resultiert bspw. die Matrix < 1 1 > für die Anfrage und für die Dokumente < 1 0 1 1 > bzw. < 1 1 0 1 >. Der Wert „1“ wird vergeben, wenn der Term im Dokument (bzw. in der Anfrage) enthalten ist, der Wert „0" drückt das Gegenteil aus.

Das zweite Dokument ist laut dieser Darstellungsform relevanter, da es beide Anfrageterme enthält. In dieser Repräsentation werden jedoch keine Aussagen über die Termhäufigkeiten miteinbezogen. Die Termhäufigkeiten werden zum einen lokal für ein Dokument berechnet. Diese wird als *term frequency* (tf) bezeichnet.

$$tf_{ij} = h(i,j)$$

Die *term frequency* gibt die Häufigkeit *h(i,j)* eines Terms t_j in einem Dokument d_i an und basiert auf der Annahme, dass ein Term den Inhalt des Dokuments umso besser repräsentiert, je öfter er in einem Dokument vorkommt.

Zum anderen werden die Termhäufigkeiten global für die gesamte Dokumentkollektion als *inverse document frequency* (idf) bestimmt.

$$idf(j) = \frac{1}{d(j)}$$

Die *inverse document frequency* basiert dagegen auf der Annahme, dass ein Term *t(j)* sich umso weniger als „Diskriminationsfaktor“ eignet, je größer die Zahl der Dokumente *d(j)* ist, in denen er vorkommt.

Die oben eingeführten Termhäufigkeiten werden für die Berechnung der Termgewichte w_{ij} herangezogen. Diese werden in der „tf-idf“-Formel als

$$w_{ij} = tf_{ij} \times idf_j$$

definiert.

Eine weitere Möglichkeit, die Ähnlichkeit zu messen, ist die Berechnung des Kosinus für den eingeschlossenen Winkel zwischen Anfragevektor $\vec{q}_j$ und Dokumentvektor $\vec{d}_j$. Der Kosinus wird mit Hilfe folgender Formel berechnet:

$$cos(\theta) = sim(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{\left|\vec{d}_j\right| \times \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,j}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$$

Das VRM ist ein anschauliches Modell zur Beschreibung der Ähnlichkeit der Anfrage in Bezug auf das Dokument, da beide in einem Vektorraum abgebildet werden. In der Abbildung 4 ist dieses Prinzip graphisch für eine dreidimensionale Dokumentenbeschreibung dargestellt. Die Dreidimensionalität wurde nur für die graphische Darstellung gewählt. Grundsätzlich können mit dem VRM n-dimensionale Dokumentbeschreibungen erstellt werden.

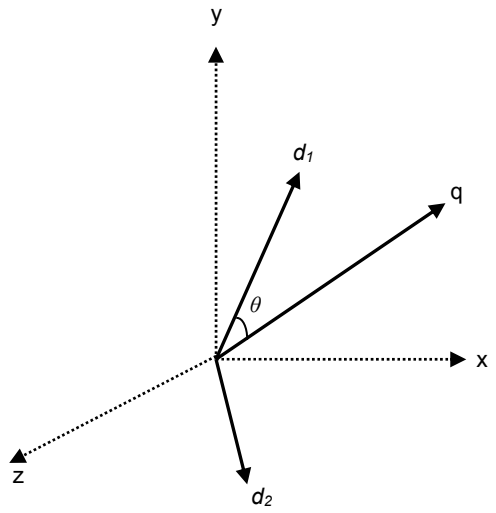

Abb. 4: *Beispiel für eine Vektorraum-Darstellung mit den Dokumenten d, der Anfrage q und dem Kosinus von θ als Ähnlichkeitsmaß für $sim(d_j, q)$.*

Ähnliche Dokumente sind im Vektorraum nahe beieinander und liegen in derselben Winkelrichtung, d.h. je kleiner der Winkel zwischen den Vektoren ist, desto ähnlicher sind die Dokumente.

Bei den Abstandsmessungen ist folgendes zu beachten:

- Ein Dokument hat zu sich selbst die Ähnlichkeit eins.
- Unvereinbare Dokumente haben die Ähnlichkeit null.
- Die Richtungen sind unbedeutend. Nur die Entfernungen zwischen Punkten geben Auskunft über die Ähnlichkeit

Das Modell hat heuristische Komponenten, woraus sich ein theoretisches Problem ergibt: es wird von eine Unabhängigkeit der Terme ausgegangen. Diese Annahme ist falsch, da z.B. “digitale Kamera” gängiger ist als bspw. “digitale Waschbretter” (vgl. Kuhlen, Griesbaum 2001, 16).

Eines der ersten Retrieval-Systeme, in denen das Vektorraummodell implementiert wurde, ist das SMART-System, das an der Cornell University in der Arbeitsgruppe von Gerard Salton entwickelt wurde (vgl. Salton, McGill 1987, 127). Für allgemeine Dokumentsammlungen liefert dieses Modell sehr gute Ergebnisse und findet deswegen wachsende Popularität bei Internetsuchmaschinen. Es ist heute die am häufigsten eingesetzte Methode des Information Retrieval (vgl. Fuhr 1997).

Das *probabilistische Modell*, das auch das *Partial-Matching* verwendet, basiert auf dem probabilistischen Ordnungsprinzip (Ranking). Die Anfrageauswertung erfolgt mit Hilfe von Wahrscheinlichkeiten, indem das System abschätzt, wie wahrscheinlich es ist, dass ein bestimmtes Dokument für eine Anfrage relevant ist (vgl. Baeza-Yates, Ribeiro-Neto 1999, 30). Diesen Wahrscheinlichkeiten entsprechend stellt das System die ermittelten Dokumente in eine Rangfolge. Die Suchergebnisse werden nach ihrer wahrscheinlichen Relevanz geordnet. Für spezifische Dokumentsammlungen liefert das Modell gute Ergebnisse. Ein Nachteil ist aber, dass die Häufigkeit eines Terms innerhalb eines Dokuments nicht beachtet wird.

2.3 Information Retrieval Techniken

Die hier beschriebenen Techniken verfolgen das Ziel möglichst viele relevante Dokumente bei gleichzeitig hoher *Precision* (wieder) zu finden. Damit Dokumente wieder gefunden werden können, müssen diese zunächst erschlossen und in eine Repräsentationsform gebracht werden.

2.3.1 Erschließung der Dokumente

Die inhaltliche Erschließung von Dokumenten, bzw. Texten, erfolgt im Wesentlichen durch drei Methoden: *Indexierung*, *Abstract-Erstellung* und *Clustering*[3].

Bei der *Indexierung* wird prinzipiell unterschieden zwischen automatischer und intellektueller Indexierung, eine Mischform beider Ansätze wird als computerunterstützte Indexierung bezeichnet (vgl. Luckhardt 1996).

Einen Text indexieren bedeutet, Begriffe für den Text zu vergeben, die dessen Textinhalt repräsentieren. Diese Begriffe werden *Deskriptoren* genannt und werden nach der Analyse der Dokumente vergeben. Die Dokumente werden aufbereitet, um effizient nach Informationen suchen zu können. Im Gegensatz zur Suche mit *regulären Ausdrücken*, wird eine Menge von Texten einmalig indexiert, um später schnell durchsucht werden zu können.

Die Deskriptoren werden in einer *Deskriptorenliste* zusammengetragen. Diese enthält formale und inhaltliche Informationen und ist das Ergebnis der Indexierung. Sie präsentiert sich als *invertierte Liste*. Die Erstellung der Deskriptorenliste wird etwas später in diesem Kapitel im Zusammenhang mit der Stoppworteliminierung erläutert.

[3] *Clustering-Verfahren* können im Unterschied zur *Indexierung* und *Abstract-Erstellung*, sowohl auf einzelne Antwortmengen, als auch auf ganze Dokumentkollektionen angewendet werden. Aus diesem Grund sind sie Bestandteil des Kapitels 2.3.2 angeführt.

Das Ziel der Indexierung ist es, relevante Dokumente zu finden, thematisch zusammengehörige Dokumente zu verknüpfen und deren Relevanz zu bestimmen. Dabei soll so erschöpfend (*exhaustivity)* und so spezifisch (*specificity*) wie möglich indexiert werden. Diese Faktoren entscheiden die Qualität des IRS und werden im Kapitel 2.5.2 näher erläutert.

Bei der *Indexierung* werden die Dokumente in eine Repräsentationsform gebracht, indem verschiedene IR-Techniken angewandt werden. Hierbei dienen die in den Dokumenten auftretenden Terme als Basis. Die Dokumentrepräsentation reduziert zum einen unterschiedliche bedeutungsgleiche Formulierungen auf die gleiche Repräsentation und steigert damit den *Recall*. Zum anderen bildet sie unklare Formulierungen, d. h. Mehrdeutigkeiten, eindeutig ab und steigert damit die *Precision*. Die wichtigsten IR-Techniken, die der inhaltlichen Erschließung der Dokumente dienen, sind:

- Stoppworteliminierung
- Grundformenreduktion (*Stemming*)
- Synonymlisten
- Thesauri und Ontologien

Für die Extraktion von Indextermen ergeben sich folgende Schritte:

1. Mittels einer *lexikalischen Analyse* werden die einzelnen Textwörter (*tokens*) identifiziert.

2. Erstellung einer *Deskriptorenliste* - Um eine Deskriptorenliste zu erstellen, wird zunächst einmal die Termfrequenz für jeden Term ermittelt, d.h. wie oft ein Begriff im Dokument vorkommt. In einem zweiten Schritt wird die Häufigkeit für jeden Begriff in der gesamten Dokumentkollektion ermittelt. Das Ergebnis ist eine Liste, die die Terme mit ihrer Häufigkeit enthält. Diese Liste wird nun nach abnehmender Häufigkeit sortiert und ein oberer Schwellenwert wird festgelegt. Oberhalb dieses Schwellenwertes befinden sich Funktionswörter mit hoher Frequenz, sog. *Stoppwörter*, die die *Stoppwortliste* (oder auch *Negativliste* genannt), darstellen. Sie beinhaltet besonders häufig verwendete Wörter (z.B. das Wort „und“), die sich

zur Unterscheidung von Dokumenten nicht eignen. Aufgrund ihrer geringen Relevanz werden *Stoppwörter* nicht indiziert und sollen nicht gesucht werden.

Dazu werden die in den *Stoppwortlisten* aufgeführten Wörter mit den in Dokumenten und Anfragen vorkommenden Wörtern verglichen. Bei Übereinstimmung werden die Wörter nicht in den Index geschrieben bzw. aus der Suchanfrage gelöscht.

Gleichermaßen wird mit den Wörtern unterhalb des festgelegten niederen Schwellenwertes verfahren. Diese Terme eignen sich aufgrund ihrer geringen Auftretenswahrscheinlichkeit nicht zur Beschreibung von Dokumenten. Diese Prozedur führt zusammen mit der Stoppworteliminierung zu einer entscheidenden Verringerung der Größe des Indizes. Die Wörter mit einer mittleren Frequenz werden in der Liste beibehalten und stellen die Deskriptoren dar. Sie haben die höchste Entscheidungsstärke um relevante Dokumente für eine Suchanfrage nachzuweisen.

Im Rahmen dieser Arbeit wurde eine Stoppwortliste für die tschechische Sprache erstellt. Auf Stoppwortlisten und deren Erstellung wird im Kapitel 5.2 genauer eingegangen.

3. Reduktionsverfahren der Terme auf einen „Grundterm“

   a) Die *Grundformenreduktion* (auch *Stemming* oder *Conflation* genannt) kann manuell oder automatisch erfolgen. Dieses sprachspezifische Verfahren überprüft die Terme auf Endungen wie z.B. Deklinations- und Konjugationsendungen, um diese ggf. zu entfernen. Auch durch das *Stemming* wird wie bei der *Stoppworteliminierung* die Anzahl der Terme im Index deutlich verringert und führt so zu einer massiven Beschleunigung des Suchprozesses. Die *Stemming-Prozedur* stellt in dieser Arbeit einen wichtigen Bestandteil dar und wird im Kapitel 5.3 ausführlich behandelt.

b) Mit Hilfe von *Synonymlisten* werden Terme auf einen Grundterm zurückgeführt. Dabei wird bspw. für die Synonyme einer Wortbedeutung[4] nur ein Term im Index festgehalten, auf den sie zurückgeführt werden. Z.B. sind die Terme „grinsen", „schmunzeln" und „lächeln" entsprechend indexiert, sodass sie auf den Term „lachen" zurückgeführt werden können.

c) Wie bei den Synonymlisten wird bei *Thesauri* und *Ontologien* jedes Wort mit seinen Synonymen unter einem gemeinsamen Begriff im Index repräsentiert. Hinzu kommt, dass sie auch Verweise auf verwandte und gegensätzliche Terme enthalten. Gleichermaßen können Wörter zu Konzepten zusammengefasst werden, die eine thematische Suche möglich machen (vgl. Sullivan 1999).

   Dadurch, dass in einem *Thesaurus* die Beziehungen zwischen den Bezeichnungen eines Fachgebiets niedergelegt sind, sollte es auch möglich sein, einzelne mehrdeutige Wörter innerhalb eines Textes eindeutig bestimmen zu können, genauer gesagt, zu determinieren, um welche spezifische Bedeutung es sich jeweils handelt und damit zu entscheiden, ob die Terme als Deskriptoren verwendet werden sollen

4. *Kompositazerlegung* und *Merhrwortgruppenanalyse* sind zwei weitere Verfahren, um Deskriptoren zu ermitteln. Im ersteren werden die Komposita in Morphemlexika zerlegt. Im zweiteren die Stammformen zu Phrasen (Mehrwortbegriffe) zusammengestellt. In den folgenden Beispielen werden deutsche Mehrwortdeskriptoren mit ihrem tschechischen Pendant angeführt:

   „Das Deutsche Rote Kreuz" - *Český Červený Kříž*
   „(deutsche) Volksbank" - *Česká Národní Banka*
   „Angewandte Sprachwissenschaft" - *Aplikovaná Lingvistika*

   Meistens sind die Mehrwortdeskriptoren bereits von einem Lexikon erfasst. Dagegen sind aber *komplexe Deskriptoren* nicht erfasst. Diese bestehen aus

[4] Der begriff Synonym ist hier als bedeutungsähnlicher Term zu verstehen.

syntaktischen Strukturen, die sinntragend in Bezug auf Deskriptoren sind. Eine intensive Auseinandersetzung mit dieser Thematik ist bei Nohr (2000)[5] zu finden.

Beim Indexierungsprozess können verschiedene Probleme auftreten, die in folgender Tabelle zusammengefasst werden:

Homographen	Wörter mit gleicher Schreibweise, aber mit unterschiedlicher Bedeutung	Wach-stube vs. Wachs-tube
Polyseme	Wörter mit mehreren Bedeutungen	Bank = Geldinstitut oder Sitzgelegenheit
Flexionsformen	entstehen durch Konjugationen und Deklinationen eines Wortes	Haus-Hauses-Häuser, schreiben-schrieb-geschrieben
Derivationsformen	Verschiedene Wortformen zu einem Wortstamm	Formatierung-Format-formatieren
Komposita	*Mehrwortdeskriptoren*[6] *komplexe Deskriptoren*[7]	Information Retrieval, retrieval of information

Tab. 1: *Probleme bei der Freitextverarbeitung*

Weiterhin sind von den informationslinguistischen Verfahren für die Indexierung folgende Aufgaben zu lösen:

Bei den Komposita kann es bei der Zerlegung in Morphemlexika zu Problemen kommen, denn nicht alle Wortbildungsmuster sind produktiv: les-bar vs. Un-kaputt-mach-bar-keit. Die Entscheidung, an welcher Stelle ein Wort zu zerlegen ist, ist nicht immer leicht zu treffen (z.B. Glücks-automaten vs. Glück-sau-tomaten). Bei der Aufteilung eines zusammengesetzten Wortes in seine Bestanteile kann in einigen Fällen eine Verschiebung der Bedeutung stattfinden, wie das Beispiel *Dachstuhl* veranschaulicht.

Die Probleme in der Kompositazerlegung treten vermehrt in der deutschen Sprache auf und sind in der tschechischen Sprache in dieser Form nicht aufzufinden. Eine

[5] http://www.iuk.hdm-stuttgart.de/nohr/KM/KmAP/Indexing.pdf
[6] auch *Mehrgliedrige Ausdrücke. Diese* müssen als solche erkannt werden, z.B.: „Das Deutsche Rote Kreuz“ - *Český Červený Kříž* und „Angewandte Sprachwissenschaft“ - *Aplikovaná Lingvistika*
[7] Diese bestehen aus syntaktischen Strukturen, die sinntragend in Bezug auf Deskriptoren sind, z.B.: Adjektiv-Substantiv-Strukturen: *blaue Himmel*

detaillierte Auseinandersetzung mit den sprachspezifischen Problemfällen für Tschechisch findet im Kapitel 3.1 statt.

Das System muss ferner in der Lage sein, ein Wort, einen Satz und eine Abkürzung als solche zu interpretieren, sowie die Funktion verschiedener Sonderzeichen erkennen. Des Weiteren muss im Rahmen der Indexierung die Schreibweise normalisiert werden und festgelegt werden, in welcher Weise die Sonderzeichen geschrieben werden.[8] Eine nicht festgelegte Schreibweise kann in manchen Fällen ein Problem darstellen kann, wie das folgende Beispiel veranschaulicht: *Masse* vs. *Maße*.

Zur Schreibweisenormalisierung zählt auch das *Case-folding*. Dabei werden alle Großbuchstaben in Kleinbuchstaben konvertiert. Dies hat den Vorteil, dass der Benutzer sich über die Groß- und Kleinschreibung eines Suchbegriffes keine Gedanken machen muss. Es gibt jedoch Anwendungsfälle, bei denen eine Unterscheidung zwischen Groß- und Kleinschreibung wünschenswert ist. Beispielsweise werden Eigennamen definitionsgemäß groß geschrieben und sollten auch in dieser Art und Weise gesucht werden können (vgl. Notes 1999). Neben der Indexierung wird auch das *Abstracting* (Abstract-Erstellung) für die inhaltliche Erschließung der Dokumente herangezogen.

Ein *abstract* ist eine Art Zusammenfassung oder Inhaltsangabe des Originaltextes. Es reduziert den Text auf die wesentlichen inhaltskennzeichnenden Elemente. Die Abstract-Erstellung kann manuell oder automatisch erfolgen. Dazu gibt es zahlreiche Verfahren, die nach dem Extraktionsprinzip funktionieren. Das Ziel ist es, zentrale Wörter zu identifizieren und Sätze mit einer gewissen Konzentration an signifikanten Termen zu extrahieren. Das abstract sollte informativ und deskriptiv sein. In ihm sollten die wichtigsten Ergebnisse, Zahlen und spezifische Informationen enthalten sein, sowie was Gegenstand der Studie ist.

Weiterhin gibt es zahlreiche Indikatoren, mit deren Hilfe die Entscheidung getroffen wird an welcher Stelle, welcher Satz, bzw. welches Wort extrahiert wird. Zu den Indikatoren zählen vorab bestimmte Schlüsselwörter und Wörter, die im Titel vorkommen. Weiterhin spielt die Gewichtung der Position eines Wortes eine

[8] Die deutschen Sonderzeichen umfassen die Umlaute und das Eszett. Die sprachspezifischen Zeichen im Tschechischen werden im Kapitel 3 angeführt.

entscheidende Rolle. So werden Wörtern/ Sätzen, die am Anfang eines Absatzes stehen, eine größere Bedeutung zugemessen als Wörtern/ Sätzen mitten im Absatz. Indikatoren sind auch sog. TOPIC-Sätze, die bspw. mit „Ich bin der Meinung, dass….“ oder „I believe…“ beginnen. Auch sog. Reizwörter deuten in bestimmten Umgebungen auf eine höhere Bedeutung der darauf folgenden Wörter („Sinn dieses Artikels ist…“ oder „wie oben erwähnt…).

2.3.2 Wiederauffinden der Dokumente

Die im vorangehenden Kapitel genannten Techniken führten Operationen innerhalb eines Dokumentes aus. Im Unterschied dazu, arbeiten die in diesem Kapitel vorgestellten Verfahren über ganze Dokumentbestände. Dies sind sehr rechenintensive Verfahren, die erst durch die technologischen Fortschritte den letzten Jahren ermöglicht wurden. Die am meisten verbreiteten Verfahren sind:

- Clustering-Verfahren
- Statistische Thesauri
- Relevance Feedback

Der engl. Begriff „clustern“ bedeutet klassifizieren, Ähnliches zusammenführen. Beim *Clustering* wird hauptsächlich die Ähnlichkeit von Dokumenten genutzt, um relevante Dokumente zu finden. Bei der Dokumentklassifikation werden thematisch ähnliche Dokumente gruppiert, bei der Termklassifikation geschieht eine Gruppierung von thematisch ähnlichen Termen. Auch die Kombination von beiden, also eine gleichzeitige Dokument- und Termklassifikation existiert.

Unter einem *Cluster* wird im Zusammenhang mit IR-Systemen eine Gruppe von Dokumenten verstanden, die untereinander eine hohe Ähnlichkeit aufweisen. Zwischen Dokumenten, die unterschiedlichen Clustern zugeordnet sind, soll die Ähnlichkeit dagegen möglichst gering sein (vgl. Frakes, Baeza-Yates 1992, 419ff).

Eine Suchanfrage wird lediglich mit den Clusterzentren (*Zentroiden*) verglichen. Sie bilden das „beste“ Element des Clusters. Die Dokumente des Clusters, die die höchste Ähnlichkeit auf eine Anfrage aufweisen, werden anschließend als Resultat

ausgegeben. Dieses Vorgehen reduziert die Anzahl der notwendigen Vergleichsoperationen und damit die Suchzeit. Die Qualität der Suchergebnisse hängt jedoch stark von der Art und Weise ab, wie die Dokumentcluster gebildet werden (vgl. Frakes, Baeza-Yates 1992, 419ff).

Die Berücksichtigung der Abhängigkeiten, im Sinne von Ähnlichkeiten der Dokumente, ist ein großer Vorteil, denn fast alle anderen IR-Modelle nehmen an, dass die Dokumente unabhängig voneinander sind. Ein Nachteil ist jedoch, dass das Clustering-Modell im Vergleich zu den anderen Verfahren eine deutlich schlechtere Retrievalqualität aufweist.

Statistische Thesauri ermöglichen dem Nutzer eine Umgebungssuche durchzuführen, indem er bspw. gezielt nach Dokumenten suchen kann, in denen die Terme dicht beieinander stehen. Dazu werden Klassen von Clustern gebildet. Damit die Klassen aussagekräftig sind, müssen sie in der richtigen *Granularität* vorliegen. Unter Granularität wird die Auflösung verstanden, mit der die Positionen der einzelnen Terme innerhalb eines Dokumentes aufgezeichnet werden. Anders ausgedrückt ist es die Genauigkeit, mit der die Position eines Terms lokalisiert werden kann. Wird beispielsweise für jedes Wort die genaue Position im Dokument im Index vermerkt, so liegt eine sehr *feine Auflösung* vor. Wählt man eine *grobe Auflösung*, so wird lediglich ein Verweis auf den jeweiligen Satz oder das jeweilige Dokument, in dem sich das Wort befindet, im Index gespeichert. Dadurch reduziert sich einerseits der Platzbedarf des Index beträchtlich, andererseits verliert man die Möglichkeit einer einfach zu realisierenden Umgebungssuche (vgl. Witten et al. 1994, 52).

Im allgemeinen ist der Recall eines IR Systems begrenzt, es werden also in den seltensten Fällen alle relevanten Dokumente für eine Suchanfrage gefunden. Somit stellt sich die Frage, wie man die ausstehenden relevanten Dokumente auffinden kann (vgl. Frakes, Baeza-Yates 1992, 241).

Relevance-Feedback soll diesem Problem Abhilfe schaffen und so die Performance des IR Systems positiv beeinflussen. Es ist eine Methode der Anfrageoptimierung, genauer gesagt, der Anfrageerweiterung. Es werden zwei Arten von Relevance-Feedback unterschieden:

- *User-Relevance-Feedback* – Hier werden dem Benutzer Begriffe aus den relevanten Ergebnissen der Anfrage zur Auswahl vorgelegt. Diese Begriffe werden folgendermaßen erzeugt: Die Suchbegriffe der zuvor als relevant eingestuften Dokumente werden zu den ursprünglichen Suchbegriffen hinzugefügt. Die Suchbegriffe, die auf nicht-relevante Dokumente verweisen, werden aus der ursprünglichen Suchanfrage entfernt. Gleichzeitig wird das Gewicht der hinzugefügten Begriffe erhöht und das der entfernten reduziert (vgl. Salton, McGill 1987, 130).
- *Blind-Relevance-Feedback* (auch *Pseudo-Relevance-Feedback*) – Im Unterschied zum User-Relevance-Feedback, bestimmt das System automatisch die Relevanz anhand vorher bestimmten Kriterien.

Die ursprünglichen Suchanfrage wird in einem nächsten Schritt reformuliert und mit den gespeicherten Dokumenten abgeglichen. Abb. 5 zeigt den Ablauf für den IR-Prozess mit Blind-Relevance-Feedback (BRF).

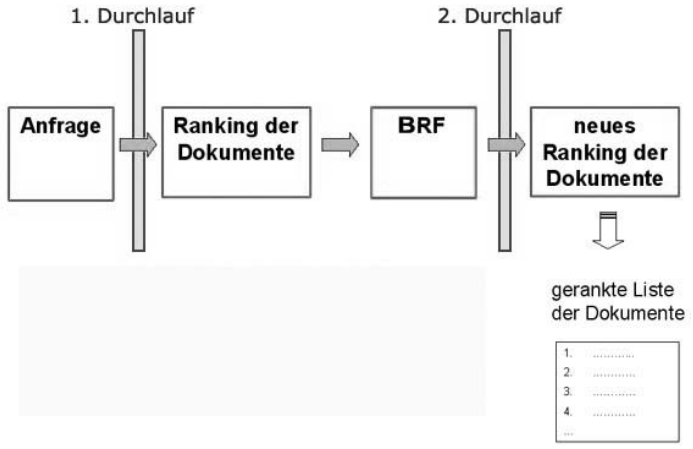

Abb. 5: *Blind-Relevance-Feedback im IR-Prozess*

Eine intensive Auseinandersetzung und eine Dokumentation erfolgreicher Retrievalergebnisse durch den Einsatz von BRF im monolingualen Zusammenhang liefern die Arbeiten von Carpineto et al. (2001) und Lam-Adesina, Jones (2001). McNamee und Mayfield (2002) untersuchten BRF im CLIR-System unter dem Gesichtpunkt des Zusammenspiels zwischen den Übersetzungsressourcen und BRF. Auch sie berichten von positiven Entwicklungen durch die Verwendung von BRF.

2.4 Evaluierung von Information Retrieval Systemen

Die informationswissenschaftliche *Evaluation (Evaluierung)* bedeutet die Bewertung, bzw. Beurteilung von Prozessen oder Ergebnissen. Bei der Beurteilung eines IRS wird laut (Korfhage 1997, 191ff.) unterschieden zwischen

- der Bewertung der Ergebnisse,
- des Information Retrieval Systems,
- der Zufriedenheit des Benutzers,
- und des Nutzens der Information

in Bezug auf die Fragestellung, auf die Frage und auf den Informationsbedarf. Für diese Bewertungskriterien werden zwei Größen herangezogen: Effizienz und Effektivität.

2.4.1 Effizienz

Das Kriterium *Effizienz* erlaubt die quantitative Beurteilung eines IRS. Unter Effizienz versteht man vor allem die Kosten und die Zeit, um eine Aufgabe zu erfüllen, d. h. Ziel ist der möglichst sparsame Umgang mit Systemressourcen für eine bestimmte Aufgabe. Zu diesen Ressourcen zählen hauptsächlich der Speicherplatz, die CPU-Zeit, die Anzahl Input-/Output-Operationen, die Antwortzeiten und der Arbeitsaufwand des Nutzers (vgl. Salton, McGill 1983, 167).

Abgesehen von diesen quantitativen Messungen kann auch eine qualitative Beurteilung von IR Systemen vorgenommen werden. Diese werden unter Effektivität zusammengefasst.

2.4.2 Effektivität

Die *Effektivität* beschreibt die Fähigkeit des Systems, dem Benutzer die benötigte Information in der für ihn best möglichsten Form anzubieten, dies bedeutet für ein IR-System, möglichst präzise und erschöpfend auf eine vom Nutzer gestellte Anfrage zu antworten. (Van Rijsbergen 1979, 145) definiert die Effektivität eines IR-Systems als

> „a measure of the ability of the system to retrieve relevant documents while at the same time holding back non-relevant ones“ .

An dieser Stelle tritt das Problem der Relevanz auf. Was zeichnet ein relevantes Dokument aus? Dieser Frage wird im Kapitel 2.4.3 nachgegangen.

Für die Evaluierung der Effefektivität von IR-Systemen haben sich zwei Standardmaße durchsetzen können: *Recall* und *Precision*. Die dazugehörigen Kriterien sind *exhaustivity* und *specificity* (vgl. Frakes, Baeza-Yates 1992, 10).

Das Kriterium *exhaustivity* (auch *Vollständigkeit* genannt) wird durch das Maß *Recall* ausgedrückt. Der *Recall* gibt also den Anteil der relevanten Dokumenten, die nachgewiesen wurden, im Verhältnis zu allen relevanten Dokumenten wieder.

Was die den Dokumentenbestand betrifft, so lässt sich dieser nach den Kriterien *relevant* und *gefunden(selektiert)* in folgende Mengen unterteilen:

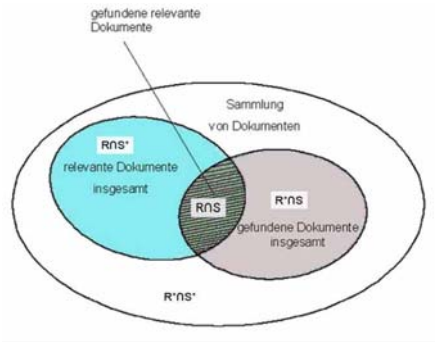

Abb. 6: *Menge der relevanten und gefundenen Dokumente*

Diese Mengen können in einer Tabelle folgendermaßen festgehalten werden:

	gefunden	nicht gefunden
relevant	A = R∩S	C = R∩S*
nicht relevant	B = R*∩S	D = R*∩S*

Tab. 2: *Die Mengen eines Dokumentbestands, die sich aus den Kriterien relevant und gefunden bilden*

A = „Treffer“ relevant gefundene Dokumente
B = „Ballast“ nicht relevante gefundene Dokumente
C = „Silence“ vermisste relevante Dokumente
D = „Umgangene Dokumente“ nicht nachgewiesene und nicht relevante Dokumente

Für den *Recall* ergibt sich daraus folgende Formel:

$$Recall = \frac{gefundene\ relevante\ Dokumente}{relevante\ Dokumente\ insgesamt} = \frac{A}{A + C}$$

Wünschenswert ist ein Wert möglichst nahe bei 1.

Das Kriterium *specificity* wird durch das Maß *Precision* ausgedrückt und betrifft die Genauigkeit des Systems durch dessen Rückhaltequote. Die *Precision* (auch *Trefferquote* genannt) gibt den Anteil der gefundenen relevanten Dokumente im Verhältnis zu der Gesamtzahl der nachgewiesenen Dokumente wieder, also wie viele relevante Dokumente sich in den insgesamt gefundenen Dokumenten befinden. Sie stellt somit ein Maß für die Güte der gefunden Dokumente dar. Auch hier ist ein Wert nahe 1 erstrebenswert.

$$Precision = \frac{gefundene\ relevante\ Dokumente}{gefundene\ Dokumente\ insgesamt} = \frac{A}{A + B}$$

Der Idealwert für Recall=1 wird erreicht, wenn C=0; für Precision=1 muss B=0 sein.
Die beiden Maße sind gegenläufig. Bei guter Vollständigkeit ergibt sich eine geringe Präzision und umgekehrt.

Der Nutzer bekommt eine nach Relevanz sortierte Liste (*Ranking-Liste*) und geht diese von oben nach unten durch, wobei sich die *Recall*- und *Precision*-Werte von

Dokument zu Dokument ändern. Die Tab. 3 zeigt exemplarisch das Ranking für die Anfrage q.

Menge der relevanten Dokumente:

$Rq=\{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$

Ranking für query q:

1. d_{123}	6. d_9	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{56}	8. d_{129}	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d_8	10. d_{25}	15. d_3

Tab. 3: *Beispiel für eine gerankte Liste*

Recall und Precision:

	Recall		Precision	
d123	10%	aller	100%	(1 von 1)
d56	20%	relevanten	66%	(2 von 3)
d9	30%	Dokumente	50%	(3 von 6)
d25	40%		40%	(4 von 10)
d3	50%		33%	(5 von 15)
-	60%		„0%“	

Tab. 4: *Beispiel für eine gerankte Liste mit den dazugehörigen Recall- und Precision-Werten* (vgl. Baeza-Yates, Ribeiro-Neto, 1999, 74).

In der Kollektion gibt es insgesamt zehn relevante Dokumente. Die nachgewiesenen relevanten Dokumente sind gelb hervorgehoben. Das erste Dokument der Liste (d_{123}) ist also relevant und stellt somit schon 10% aller relevanten Dokumente im Korpus dar, dies bedeutet, dass es eine *Precision* von 100% und einen *Recall* von 10 % aufweist.

Um die Entwicklung der Werte besser darstellen und untersuchen zu können, werden sie in einer *Recall-Precision-Kurve* festgehalten. Die Kennwerte und zugehörigen Kurven werden mit Hilfe von Testdatenbanken und genau festgelegten Anfragen, für die zuvor eine Relevanzbeurteilung vorgenommen wurde, experimentell ermittelt. Die beiden Maße werden dokumentweise erarbeitet, d.h. nach jedem Dokument werden *Recall* und *Precision* gemessen. Dabei wird folgendermassen verfahren: Die beiden

Maße werden gegeneinander in Graphen aufgetragen, wobei den elf *Recall*-Stufen (Standardmesspunkte) zwischen 0.0 und 1.0 ihre *Precision*-Werte zugeordnet werden. Die Punkte werden durch Geraden verbunden (vgl. Abb. 7).

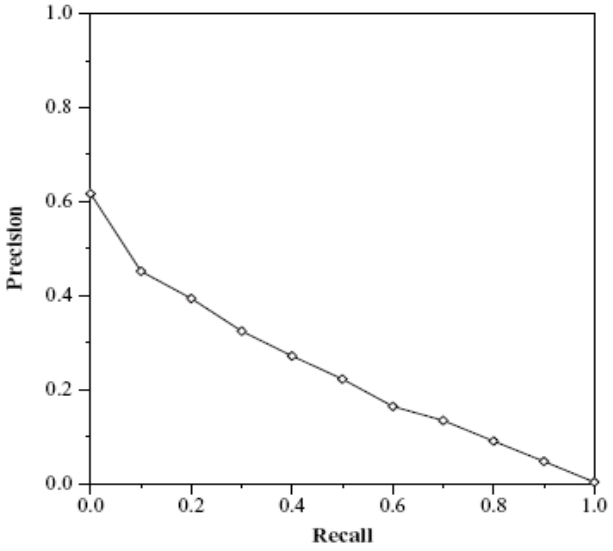

Abb. 7: *Beispiel für einen Recall-Precision-Graphen*

Die Geraden haben keine interpolierende Bedeutung, das bedeutet, dass zwischen den 11 Standardmesspunkten keine Werte definiert sind. Aus diesem Grund wird eine Interpolationsprozedur angewendet (vgl. Baeza-Yates, Ribeiro-Neto 1999, 76).

Der Vergleich von verschiedenen Systemen geschieht über den Kurvenverlauf der Systeme, dessen Koordinaten, sich aus dem Wertepaar r_i und p_i zusammensetzen und unter gleichen Bedingungen zustande kamen. Die darüber liegende Kurve wird dabei als besser eingestuft.

Es gilt:

Das Paar (r_1, p_1) ist mindestens dann besser als (r_2, p_2), wenn

$$r_1 \geq r_2 \wedge p_1 > p_2 \quad \text{oder} \quad r_1 > r_2 \wedge p_1 \geq p_2$$ ist.

Ein System ist besser als ein anderes, wenn für das eine System sowohl der Precision-Wert als auch der Recall-Wert besser ist als bei dem anderen System. Ist bei einem

System z.B. die Precision besser, dafür aber der Recall schlechter, so eignen sich die Systeme zwar für unterschiedliche Aufgaben. Die allgemeine Aussage aber, welches besser ist, kann nicht geäußert werden. (vgl. Ferber 2003).

Die Standardmaße Recall und Precision haben sich für die Evaluierung von IR-Systemen zwar etabliert, dennoch existieren einige Kritikpunkte.

Der *Recall* bezieht die *Ballast-Quote* nicht mit ein. Weiterhin wird in der Formel mit einem Schätzwert für die „Silence-Quote“ im Nenner gearbeitet. Auch bei der Berechnung der Standardmaße können sich Fehler ergeben, wenn für eine Suchanfrage keine relevante Dokumente vorhanden sind oder keine nachgewiesen werden. Um diesen Fall auffangen zu können, wurde eine weitere Größe eingeführt: die “Fallout-Quote“. Sie gibt die Retrievalleistung in Abhängigkeit von den irrelevanten Dokumenten wieder. Die „Fallout-Quote“ wird folgendermaßen definiert:

$$Fallout = \frac{nicht\ relevante\ gefundene\ Dokumente}{alle\ nicht\ relevante\ Dokumente} = \frac{B}{B+C}$$

Ein weiterer entscheidender Kritikpunkt stellt die Tatsache dar, dass angenommen wird, dass alle relevanten Dokumente in der Dokumentenkollektion bekannt sind. Bei großen Kollektionen kann oft der maximale *Recall* nicht festgestellt werden. Eine mögliche Lösung, die auch im Rahmen des Cross-Language Evaluation Forum eingesetzt wird, bietet die *Pooling-Methode* und die *Juroren*. Diese wird im Kapitel 2.7 näher erläutert.

2.4.3 Relevanz

Das bereits erwähnte Problem der Relevanz äußert sich dadurch, dass schwer zu bestimmen ist, welche Antwort die richtige für eine Anfrage an das System ist. (Bekavac 2001, 13) definiert die Relevanzbeurteilung als

> „Grad der Übereinstimmung der inhaltlichen Aussage eines Dokumentes aus der Treffermenge und der Suchanfrage“.

Je nach Perspektive kann von der System- bzw. Benutzerrelevanz gesprochen werden. Die Systemrelevanz beschreibt den Grad der formalen Übereinstimmung, die durch das System determiniert wird. Im Falle der Benutzerperspektive urteilt der Nutzer, ob das Rechercheergebnis seinem Informationsbedarf entspricht oder nicht. Dokument kann für verschiedene Anwender von unterschiedlicher Relevanz sein. Somit ist die Relevanz ein subjektives und unscharfes Beurteilungsmaß.

2.5 Multilinguales Information Retrieval

Im Unterschied zu den *monolingualen* IR-Systemen, die auf eine Anfrage in einer bestimmten Sprache nur Ergebnisse in derselben Sprache liefern, werden in *multilingualen* IR-Systemen verschiedene Sprachen berücksichtigt. Bei multilingualem IR unterstützt das System die Anfrage, sowie die Repräsentation der Ergebnisse aus der Dokumentensammlung in mehreren Sprachen. Allerdings entspricht die Sprache der Anfrage immer nur der Sprache der gefundenen Dokumente. Eine sprachenübergreifende Suche findet hier nicht statt.

Dieser sprachenübergreifende Aspekt findet sich im *cross-lingualen* Information Retrieval wider. Cross-linguale Information-Retrieval-Systeme (CLIR-Systeme) suchen relevante Informationen in einer Dokumentenkollektion, wobei die Anfrage in einer anderen Sprache formuliert wird, als die in der die Dokumente verfasst sind. Grefenstette (1998, 2) beschreibt die Aufgabe des CLIR als

> „a task of filtering, selecting and ranking documents that might be relevant to a query expressed in a different language."

In der Literatur herrscht oft keine einheitliche und eindeutige Verwendung der Terme. So werden bspw. auch die Begriffe *translingual* und *multilingual* als Synonyme für das hier beschriebene *cross-lingual* verwendet (vgl Oard 1997). Die eben eingeführte Unterscheidung zwischen monolingualem, multilingualem und cross-lingualem IR soll für diese Arbeit gelten, wobei cross-linguales IR als Untergruppe des multilingualen

IR betrachtet wird. Zusammengefasst kann gesagt werden, dass das cross-linguale IR eine Erweiterungsform des multilingualen IR darstellt.

Die Notwendigkeit von cross-lingualen Verfahren ist in der Tatsache begründet, dass im Internet der Zuwachs von nicht-englischsprachigen Dokumenten erheblich größer ist als der Zuwachs von englischsprachigen Dokumenten.

Die steigende Entwicklung der nicht-englischsprachigen Internetnutzung für den Zeitraum 1996 bis 2005 kann der Graphik Abb. 8 entnommen werden.

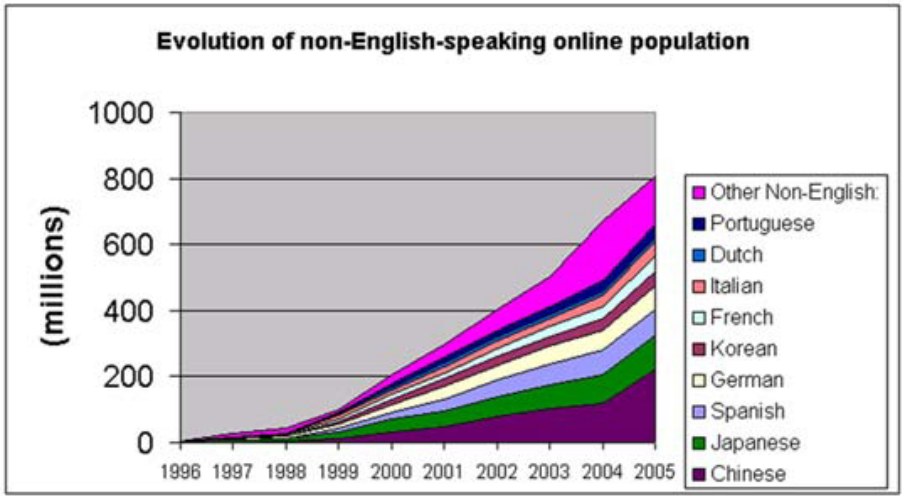

Abb. 8: *Entwicklung der nicht-sprachigen Internetnutzern*[9]

Laut den Internetstatistiken von *Global Reach*[10] waren im September 2004 weltweit 64.8% der Internetnutzer nicht englischsprachig. Der englischsprachige Teil stellte 35.2% dar. Ein entscheidender Aspekt stellt gewiss die Tatsache dar, dass nicht alle Benutzer korrekte englische Anfragen formulieren können.

Des Weiteren besteht die Annahme, dass häufig die Fremdsprachenkenntnisse eines Benutzers zwar so weit gehen, dass sie ausreichen, um die Relevanz von Dokumenten in dieser Sprache abschätzen zu können. Nichtsdestotrotz bleiben aber die Probleme bei der Formulierung von Anfragen bestehen. Das CLIR-System schafft dem Benutzer

[9] http://global-reach.biz/globstats/evol.html
Die Angaben zu der "Online-Bevölkerung" geben nicht die Anzahl der Personen wider, die die betreffende Sprache sprechen, sondern beziehen sich auf die Anzahl der Personen im Internet für die jeweilige Sprache, z.B. Muttersprachler
[10] http://global-reach.biz/globstats/index.php3

die Möglichkeit, die Anfrage in seiner Muttersprache zu stellen und auf Dokumente zuzugreifen, die in fremden Sprachen vorliegen. Damit hat er Zugriff auf alle vorliegenden relevanten Dokumente, unabhängig von der Sprache, in der die Anfrage erfolgte.

Die wesentliche Aufgabe des IR, nämlich die Suche nach relevanten Dokumenten, bleibt im multilingualen IR bestehen. Diese Aufgabe wird im multilingualen Zusammenhang um die Sprachkomponente erweitert.[11] Bei cross-lingualem IR entsteht die Notwendigkeit die Anfrage umzuformen oder das Dokument oder auch beides, um sie in eine gemeinsame terminologische Repräsentation zu bringen. Laut Grefenstette (1997, 3) entstehen dadurch drei Hauptprobleme beim CLIR:

- die Übersetzung des Terms an sich (also wie ein Term in einer anderen Sprache korrekt ausgedrückt wird)

- die Auswahl der möglichen Übersetzungsvarianten für einen Term

- die Gewichtung der Übersetzungsvarianten, falls mehr als ein Term ausgewählt wurde. Dabei ist zu beachten, dass bei der Verwendung von mehreren Übersetzungsalternativen für einen Term, der Recall erhöht wird. Weiterhin ist anzumerken, dass ein Term mit mehreren Übersetzungsvarianten für einen Term kein größeres Gewicht erhalten darf als ein Term mit nur einer Übersetzungsvariante.

Die ersten zwei Probleme sind in den Bereichen der maschinellen Übersetzung angesiedelt. Das dritte Problem findet sich im Bereich der Ergebnisliste. Die Überwindung dieser Sprachbarrieren geschieht durch die Verwendung von Übersetzungsressourcen. Diese beruhen entweder auf korpusbasierten oder wissensbasierten Ansätzen. Beim ersteren werden durch die Verwendung von parallelen oder vergleichbaren Dokumentenmengen assoziative Ähnlichkeitsthesauri gebildet. Diese werden für die Übersetzung herangezogen. Beim zweiteren werden Wörterbücher und Ontologien benutzt. Einen guten Überblick über die eben genannten Ressourcen geben Baeza-Yates und Ribeiro-Neto (1999, 143).

[11] Da das CLIR als Untergruppe des multilingualen IR angesehen wird, erbt das CLIR die Eigenschaften des multilingualen IR.

Unabhängig davon, welche Ressource verwendet wird, entstehen grundsätzlich folgende Problemfälle:

- nicht übersetzte Terme, da sie von der Ressource nicht erfasst wurden
- allgemeine Übersetzungsprobleme (z.B. : Ambiguität der Terme, Komposita,...)
- Anfragen sind zu kurz (Terme, die in relevanten Dokumenten vorkommen, fehlen in der Anfrage)
- Das Vorhandensein von Eigennamen und deren Erkennung als solche. Mit dieser Problematik haben sich Mandl und Womser-Hacker (2003) intensiv auseinandergesetzt.

Das bereits genannte Problem der Gewichtung der Übersetzungsvarianten gestaltet sich schwieriger, falls eine mehrsprachige Ergebnisliste erstellt wird. Das Zusammenführen von mehreren Trefferlisten in eine einzige wird als *Fusion* bezeichnet. Diese Liste enthält die Ergebnisse aus verschiedensprachigen Kollektionen, wobei das Problem zu beachten ist, dass möglicherweise unterschiedliche Gewichtungsalgorithmen verwendet werden. In diesem Fall sollte diese entsprechende behandelt werden.

Im Rahmen der multilingualen Retrievalexperimente der Universität Hildesheim wurde der Fusion-Ansatz durchgeführt und in Hackl (2004, 57) beschrieben.

2.6 Das MIMOR-Modell[12]

Die Idee für das MIMOR-Modell entstand bei der Beobachtung der TREC[13]-Ergebnisse. Auffallend war, dass unterschiedliche IR-Systeme zwar vergleichbar gute Precision-Werte erzielten, trotzdem aber nicht zu den gleichen Dokumenten führten (vgl. Mandl, Womser-Hacker 2000, 4ff). Die so entstandene Schnittmenge der Treffer

[12] MIMOR = Mehrfachindexierung zur dynamischen Methoden-Objekt-Relationierung im IR
[13] TREC (Text Retrieval Conference) http://trec.nist.gov/

fiel meistens relativ gering aus. Um die Ergebnisse der verschiedenen IR-Systeme in einem Gesamtergebnis zu vereinen und somit eine höhere Qualität zu erzielen, werden diese in Fusionsverfahren kombiniert.

Das Ziel MIMORs ist es, die Faktoren, die zu einem bestimmten Retrievalergebnis geführt haben, zu determinieren, um anschließend diese in verschiedene Techniken des adaptiven Modells zu integrieren, sodass eine größere Anzahl an relevanten Dokumenten gefunden wird.

Aus diesem Grund ist MIMOR

> „als Testmodul konzipiert, das in den Retrievalprozess eingeschoben wird und die Korrelation zwischen Deskribierungsmethoden und Objekteigenschaften ermittelt“ (Womser-Hacker, 1996, 284).

Das MIMOR-Modell integriert Relevance-Feedback und den Fusionsansatz in einem adaptiven Modell. Anhand von Relevance Feedback lernt das Modell, die Einzelergebnisse zu fusionieren. Der Benutzer entscheidet an dieser Stelle, welche Dokumente im Ergebnis besonders relevant sind. Durch die Speicherung und Analyse dieser Urteile lernt MIMOR, welche die besten Verfahren in konkreten Anwendungsfällen für ein Informationsbedürfnis sind. Anschließend versucht es die Zusammenhänge zwischen den angewendeten Verfahren und den Objekteigenschaften zu erfassen.

MIMOR ist durch das ihm zugrunde liegende adaptierbare Lernkonzept so angelegt, dass sich während des Einsatzes des IR-Systems die besten Verfahren durchsetzen. Zu Beginn werden alle Verfahren gleich gewichtet. Im Lauf des Lernprozesses werden die Gewichte so angepasst, dass Verfahren, die das Ergebnis positiv beeinflussen, stärker gewichtet werden.

2.7 Das Cross Language Evaluation Forum (CLEF)

Das Cross Language Evaluation Forum (CLEF) ist eine europäische Initiative zur Evaluierung der Multilingualität und Multimodalität von IR-Systemen, die aus dem *Cross Language Track* der Text Retrieval Conference (TREC) in den USA hervorgegangen ist. Die Evaluierungskampagnen finden seit 2000 jährlich statt.

CLEF wird vom *Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche* in Pisa koordiniert. Eine Auflistung der Institutionen, die zu der Organisation der verschiedenen *Tracks*[14] von CLEF 20005 beitragen, ist auf der Homepage von CLEF unter http://www.clef-campaign.org/ zu finden.

CLEF hat sich zur Aufgabe gemacht, CLIR-Systeme in einer entsprechenden Evaluierungsumgebung bezüglich ihrer Anwendung auf europäische Sprachen zu testen und ihre Leistung zu bewerten. Hierbei sollen nicht nur die größten europäischen Sprachen zum Einsatz kommen, sondern vermehrt auch die „kleinen" Sprachen Europas. Zu diesen Sprachen zählt bspw. Tschechisch, das in diesem Zusammenhang Gegenstand der vorliegenden Arbeit ist. Dabei geht es darum, Ressourcen für die tschechische Sprache im Hinblick auf eine spätere Implementierung in CLEF zu erstellen, aufzuzeigen und zu evaluieren.

Den Teilnehmern, Universitäten und andere Forschungseinrichtungen, bietet CLEF die Möglichkeit, ihre IR-Systeme an eigens entwickelten, großen Dokumentkollektionen zu testen und zu vergleichen. Diese bestehen aus den Jahrgängen 1994 und 1995 von national erscheinenden Zeitungen im SGML- oder XML-Format für die Sprachen Holländisch, Englisch, Finnisch, Französisch, Deutsch, Italienisch, Portugiesisch, Russisch, Spanisch und Schwedisch. Da die Teilnahme von Bulgarisch und Ungarisch später erfolgte, stammen die dazugehörigen Zeitungen aus dem Jahr 2002.

In dieser Arbeit sind ferner zwei weitere Kollektionen von Interesse: zum einen EuroGOV, die im Rahmen von WebCLEF verwendet wird und für die Erstellung des in Kapitel 5.4 beschriebenen Text-Katalogs für Tschechisch in dieser Arbeit herangezogen wurde und zum andern die MALACH-Kollektion, die für das Cross-

[14] *Tracks* bezeichnen die Aufgabenstellungen in CLEF.

Language Spoken Document Retrieval (CL-SR)[15] eingesetzt wird und das einzige Track ist, an dem Tschechisch teilnimmt.

Bei den eingesandten Ergebnissen erfolgt die Ermittlung der relevanten Dokumente nach der sog. *Pooling-Methode*. Dieses Verfahren wird herangezogen, wenn sehr große Testkollektionen verglichen und evaluiert werden sollen, wie auch bei TREC (vgl. Grefenstette 1998, 139). Die hinsichtlich der Relevanz am höchsten eingestuften Dokumente aller beteiligten Testsysteme werden in Ergebnislisten vereint und von den Juroren auf Relevanz bewertet. Die Unterscheidung zwischen relevanten und nicht-relevanten Dokumente ist nicht immer klar zu treffen und setzt fundierte Kenntnisse der Juroren voraus. Dabei wird laut (Baeza-Yates und Ribeiro-Neto 1999, 89) von zwei Annahmen ausgegangen:

1. Der Großteil der relevanten Dokumente befindet sich im zusammengestellten *Pool*
2. Die Dokumente, die sich nicht im Pool befinden, können als nicht-relevant eingestuft werden.

Der Fachbereich Angewandte Informationswissenschaft der Universität Hildesheim nimmt an CLEF mit einem eigenen Fusionssystem teil und beteiligt sich an der Organisation von CLEF und dem Aufbau der mehrsprachigen Anfragekorpora. [16]

[15] Weitere Informationen zu CL-SR sind unter http://clef-clsr.umiacs.umd.edu/index.html zu finden.
[16] Weiterführende Informationen hierzu unter: http://www.uni-hildesheim.de/~mandl/Forschung/Clef/index.html

Kapitel 3

Die tschechische Sprache

Tschechisch ist die Amtssprache der Tschechischen Republik und zählt seit dem 1. Mai 2004 zu den Amtssprachen der EU. Etwa 12 Millionen Menschen sprechen Tschechisch als Muttersprache (Stand 1999), von denen ca. 10 Millionen in der Tschechischen Republik leben. Gesprochen wird Tschechisch auch in den angrenzenden Ländern, vor allem in der Slowakei (50-70 Tsd. Sprecher, überwiegend tschechisch-slowakische Familien aus den Zeiten der Föderation).[17] Auch in Übersee gibt es tschechischsprachige Minderheiten. Die größte davon in Nordamerika mit 500-600 Tausend Sprechern.[18]

Die tschechische Sprache gehört, neben z.B. Polnisch, Slowakisch und Sorbisch zur westslawischen Sprachfamilie. Sie bildet zusammen mit den ostslawischen und südslawischen Sprachen den slawischen Sprachenzweig.[19] In Abb. 9 wird eine Übersicht zu den bestehenden Verwandtschaftsgraden zwischen den slawischen Sprachen gegeben.

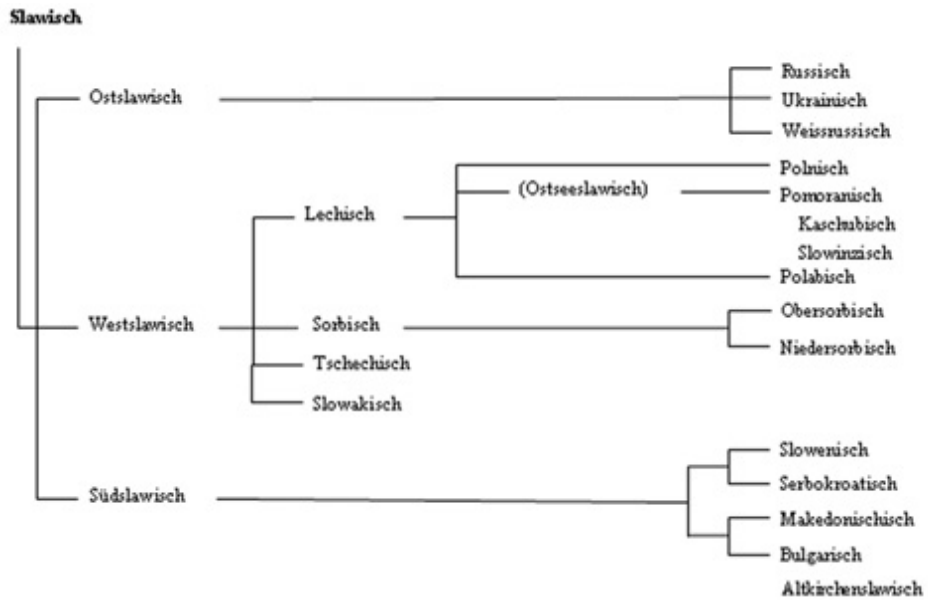

Abb. 9: *Der slawische Sprachenzweig*

[17] wikipedia
[18] http://www.czech-language.cz/overview/territories.html
[19] Janich, Nina; Greule, Albrecht (Hrsg.) (2002) *Sprachkulturen in Europa - ein internationales Handbuch*. Tübingen:Narr, S. 302.

Da die slawischen Sprachen zu der indoeuropäischen Sprachenfamilie zählen, weisen sie Gemeinsamkeiten mit bspw. dem Englischen und dem Deutschen auf. Diese Gemeinsamkeiten sind für die Informationslinguistik für die Generierung von allgemeinen Regeln von großem Interesse. In dieser Arbeit werden diese Gemeinsamkeiten bspw. für die Erstellung von Stoppwortlisten herangezogen (vgl. Kapitel 5.2). Tschechisch, Slowakisch und Sorbisch sind sich sehr ähnlich, so dass sie gegenseitig gut verständlich sind.

Seit dem 13. Jahrhundert wird das Tschechische mit lateinischen Buchstaben geschrieben. Unter dem Einfluss des *Hussitentums* war Tschechisch einige Zeit auch internationale literarische Schriftsprache in Polen und Ungarn. Das Hussitentum ist die Zeit verschiedener reformatorischer beziehungsweise revolutionärer Bewegungen im Böhmen des 15. Jahrhunderts. [20] Der Name geht auf den tschechischen Theologen und Reformator Jan Hus (* um 1370) zurück. Er führte 1406 mit seinem Werk „Orthographia Bohemica“ die diakritischen Zeichen ein und vereinfachte die Grammatik der tschechischen Sprache. Ein jüngeres Ereignis, das die tschechische Sprache beeinflusste, war die Rechtschreibreform im Jahre 1994.

Der Grundwortschatz des Tschechischen ist slawisch. Durch die lange Zugehörigkeit Tschechiens zum Deutschen Reich, bzw. zu Österreich, gibt es einen deutschen Einfluss des Deutschen auf das Tschechische. Dieser Einfluss findet sich v.a. in der Umgangssprache wider. So sind bspw. „švindlovat“, „vartovat“ und „maršírovat“ die tschechischen umgangssprachlichen Entsprechungen für die sehr ähnlich klingenden deutschen Verben „schwindeln“, „warten“ und „marschieren“. Auch kleine Füllworte wie „Au", „Pšt“ und „Nojo“ ähneln sehr den deutschen Entsprechungen. Daneben lassen sich einige französische und, v. a. aus der sozialistischen Zeit stammende, russische Wörter finden. Gegenwärtig kommt es durch die technische Entwicklung und die Internationalisierung zu einem verstärkten Eintrag englischer Begriffe.

[20] vgl. http://de.wikipedia.org/wiki/Hussiten

3.1 Die Besonderheiten der tschechischen Sprache

Damit die Leistung von CLIR-Systemen optimiert werden kann, müssen die im Prozess verwendeten Ressourcen auf die spezifischen Merkmale der beteiligten Sprachen abgestimmt werden. Unterschiedliche Sprachen verursachen mit ihren charakteristischen Eigenschaften im Verlauf des CLIR-Prozesses verschiedenartige Probleme, auf die mit bestimmten Methoden und Werkzeugen eingegangen werden sollte (vgl. Peters 2001, 4).

Um die für eine Sprache geeigneten Werkzeuge zu finden, müssen zunächst die Besonderheiten der Sprache determiniert und anschließend im Hinblick auf die Verarbeitung durch die jeweiligen Ressourcen analysiert werden. Im Folgenden werden die im Rahmen dieser Arbeit ermittelten Eigenschaften der tschechischen Sprache in einem ersten Schritt vorgestellt und in einem zweiten hinsichtlich möglicher Probleme im IR-Prozess untersucht. Eine tiefere Auseinandersetzung mit dieser Problematik erfolgt in dem Kapitel 5.

Im Gegensatz zu den meisten slawischen Sprachen wird das Tschechische nicht in kyrillischer Schrift, sondern mit lateinischen Buchstaben geschrieben. Da Tschechisch, wie alle slawischen Sprachen, über wesentlich mehr Laute verfügt, als das lateinische Alphabet Buchstaben aufweist, werden diakritische Zeichen verwendet. Laut Bußmann (1990, 177) sind diese

> „Zusätze an oder in Schriftzeichen, mit denen bestimmte Unterscheidungen getroffen werden sollen."

In erster Linie dienen diakritische Zeichen der Kennzeichnung einer unterschiedlichen Aussprache und nehmen im Tschechischen die Form von Akzenten, Häkchen (Háček) und Ringel (kroužek) an.[21]

Folgende Kodierungen sind für den tschechischen Zeichensatz geeignet: ISO 8859-2, CENTEURO, CP1250 und CP852.[22]

[21] http://www.sochorek.cz/archiv/sprachen/tschechisch/fakten.htm
[22] Institute of the Estonian Language, http://www.eki.ee/letter/chardata.cgi?lang=cs+Czech&script=latin

Weiterhin auffällig ist ein hoher Konsonantenanteil in den tschechischen Wörtern, sodass einige Wörter keine Vokale aufweisen, wie z.B. das Wort „*prst*“ (dt. „Finger“).

Die Wortstellung im Satz ist relativ frei und wird als stilistisches Mittel verwendet. Eine weitere Eigenschaft der tschechischen Sprache besteht darin, dass kaum Komposita gebildet werden.[23] Die Anzahl der Komposita ist so gering, dass sie sich in einer relativ kurzen Liste aufzählen lassen.[24]

Somit nimmt die Problematik der Kompositazerlegung im IR, die etwa vermehrt in der deutschen Sprache anzutreffen ist, nur geringe Ausmaße an. Die Kompositazerlegung im IR-Prozess könnte folglich für Tschechisch ganz vernachlässigt werden. Auf diesen Aspekt wurde bereits im Kapitel 2.3.1 genauer eingegangen.

Folgende Beispiele veranschaulichen, in welcher Form deutsche Komposita im Tschechischen aufgelöst werden:

a) Meistens wird ein Adjektiv in Kombination mit einem Substantiv verwendet:

„dezinfekcní prostredky“	*Desinfektionsmittel*
„informační materiál“	*Informationsmaterial*
„webové stránky“	*Webseiten*

b) In einigen Fällen werden deutsche Komposita auch in zwei Substantive und einer Präposition aufgeteilt:

„návod k použití“	*Gebrauchsanweisung*
„Anweisung“ + „zum“ + „Gebrauch“	

[23] Mit *Kompositum* ist hier ein einzelnes Wort gemeint, dass sich aus mehreren Wörtern zusammensetzt, im Unterschied zu Mehrwortgruppen. In welcher Form Komposita im Tschechischen auftreten, wird im folgenden Abschnitt anhand der Beispiele a bis c veranschaulicht.

[24] Eine komplette Auflistung der tschechischen Komposita befindet sich in: (Dušan 1999, 106-124). Allerdings ist zu beachten, dass in dieser Liste auch zahlreiche Eigennamen, wie bspw. das tschechische Pendant zu dem deutschen „Gottfried“, als Komposita angeführt werden.

c) Zu den vereinzelten Ausnahmen im Tschechischen für Komposita, im Sinne von einem Wort, das sich aus mehreren zusammensetzt, gehören:

„samoobsluha“ *Selbsbedienungsladen*
(samo = „selbst“) + (obsluha = „Bedienung“)

„vodovod“ *Wasserleitung*
(voda = „Wasser“) + (vedení = „Leitung“)

„trestuhodný“ *rügenswert*
(trest = „Strafe“) + (hodný = „wert/würdig“)

Tschechisch ist eine stark flektierende Sprache[25], das heißt, die grammatikalische Funktion der Wörter wird erst durch die Flexion vollständig beschrieben. Weitere Beispiele für stark flektierende Sprachen sind Arabisch und Russisch.

Die Deklination und Konjugation erfolgt bei flektierenden Sprachen mittels Endungen und/oder kleinen Änderungen im Stamm. Diese Änderungen im Stamm werden *Fusion* genannt und entstehen dadurch, dass die Morpheme Nachbarmorpheme beeinflussen und ebenfalls durch sie beeinflusst werden. Die Bildung der Deklinations- und Konjugationsendungen ist sehr vielfältig und nicht immer regelmäßig. Somit ist sie sehr schwer zu formalisieren. Laut (Bußmann 1990, 244) ist bei flektierenden Sprachen

> „eine (hinsichtlich Form und Funktion) eindeutige Segmentierung von Wurzel- und Wortbildungsmorphemen nicht möglich“.

Diese Tatsache erschwert die formale Erfassung der sprachlichen Phänomene in der tschechischen Sprache. Die formale Erfassung wird weiterhin durch die Deklination von Substantiven, Adjektiven, Pronomina und Numeralia verkompliziert. Auch Eigennamen werden dekliniert.

[25] Der Terminus *flektierende Sprachen* wird auch häufig als Synonym für *synthetische Sprachen* gebraucht

Von den ursprünglich acht indoeuropäischen Fällen der Substantive hat das Tschechische noch sieben erhalten: Nominativ, Genitiv, Dativ, Akkusativ, Lokativ, Instrumental und Vokativ.[26] Die Kasusendungen für Lokativ und Instrumental entsprechen im Deutschen den Präpositionen „von“ bzw. „mit“. Tab. 5 soll einen Überblick über die vielfältigen Deklinationen der Ausdrücke

ten mladý pán	„der junge Mann“,
ta mladá žena	„die junge Frau“
und *to nové město*	„die neue Stadt“

geben.

DEKLINATION SINGULAR					**DEKLINATION PLURAL**				
	MASKULINUM					**MASKULINUM**			
N	ten	mladý	pán	der junge Mann	**N**	ti	mladí	páni	die junge Männer
G	toho	mlad**ého**	pán**a**	des jungen Mannes	**G**	těch	mlad**ých**	pán**ů**	der jungen Männer
D	tomu	mlad**ému**	pán**ovi**	dem jungen Mann	**D**	těm	mlad**ým**	pán**ům**	den jungen Männern
Ak	toho	mlad**ého**	pán**a**	den jungen Mann	**Ak**	ty	mlad**é**	pán**y**	die jungen Männer
L	o tom	mlad**ém**	pán**ovi**	von dem jungen Mann	**L**	o těch	mlad**ých**	pán**ech**	von den jungen Männern
I	tím	mlad**ý**	pán**em**	mit dem jungen Mann	**I**	s těmi	mladými	pán**y**	mit dem jungen Männern
	FEMININUM					**FEMININUM**			
N	ta	mladá	žena	die junge Frau	**N**	ty	mlad**é**	žen**y**	die jungen Frauen
G	té	mlad**é**	žen**y**	der jungen Frau	**G**	těch	mlad**ých**	žen	der jungen Frauen
D	té	mlad**é**	žen**ě**	der jungen Frau	**D**	těm	mlad**ým**	žen**ám**	den jungen Frauen
Ak	tu	mlad**ou**	žen**u**	die junge Frau	**Ak**	ty	mlad**é**	žen**y**	die jungen Frauen
L	o té	o mlad**é**	žen**ě**	von der jungen Frau	**L**	o těch	mlad**ých**	žen**ách**	von den jungen Frauen
I	tou	mlad**ou**	žen**ou**	mit der jungen Frau	**I**	s těmi	mlad**ými**	žen**ami**	mit den jungen Frauen
	NEUTRUM					**NEUTRUM**			
N	to	nové	město	die neue Stadt	**N**	ty	nov**á**	měst**a**	die neuen Städte
G	toho	nov**ého**	měst**a**	der neuen Stadt	**G**	těch	nov**ých**	měst	der neuen Städte
D	tomu	nov**ému**	měst**u**	der neuen Stadt	**D**	těm	nov**ým**	městům	den neuen Städten
Ak	to	nové	město	die neuen Stadt	**Ak**	ta	nov**á**	měst**a**	die neuen Städte
L	o tom	o nov**ém**	měst**u/e**	von der neuen Stadt	**L**	o těch	o nov**ých**	měst**ech**	von den neuen Städten
I	tím	nov**ým**	měst**em**	mit der neuen Stadt	**I**	s těmi	nov**ými**	měst**y**	mit den neuen Städten

Tab.5 : *Deklinationstabelle für die vier tschechischen Ausdrücke: ten mladý pán -„der junge Mann“, ta mladá žena - „die junge Frau“ und to nové město -„die neue Stadt“.*

[26] Im Deutschen hingegen gibt es nur vier und im Englischen nur zwei Fälle.

Der Vokativ stellt laut Čechová, Trabelsiová und Putz (1996, 23) eine spezielle Form für die Anrede dar und kann aus diesem Grund in der Tab. 5 nicht mit angeführt werden. Ein Beispiel für den Vokativ-Gebrauch wäre:

Pane doktore! „Herr Doktor!“

Im Vergleich hierzu die Nominativ-Form: *pan doktor.*

Neben den zahlreichen Kasusendungen existieren im Tschechischen bei den Substantiven drei Genera: maskulinum, femininum und neutrum. Ferner existieren zwei Pluralformen für die Substantive, die je nach Anzahl verwendet werden. Nach der Mengenangabe zwei, drei oder vier erfolgt in der Pluralform der *Nominativ Plural*. Sollte das betreffende Substantiv in einer größeren Anzahl als vier vorkommen (also fünf oder mehr), so wird in der Pluralform der *Genitiv Plural* angewendet. Diese Tatsache hat eine große Menge verschiedener Wortformen zur Folge.

Im *Akkusativ Singular* und *Nominativ Plural* der Maskulina gibt es unterschiedliche Formen für belebte (Belebtheitskategorie) und unbelebte Wesen. Des Weiteren ist auch die Konjugation der Verben stark ausgeprägt, wobei es die drei Zeiten Vergangenheit, Präsens und Futur gibt. Neben den drei Modi Indikativ, Imperativ und Konjunktiv existieren von den meisten Verben zwei Aspekte, den vollendeten und den unvollendeten. Das Verb verfügt über die Kategorien von Aspekt (perfektiv und imperfektiv) und Tempus (Präsens, Futur, Präteritum), Person, Numerus und Modus (Imperativ, Konditional) (vgl. Čechová, Trabelsiová, Putz 1996, 23).

Ferner existieren im Tschechischen zwei bedeutungsunterscheidende Formen des Passivs: das *Vorgangspassiv* und das *Zustandspassiv*. Folgendes Beispiel soll den Unterschied verdeutlichen:

okno bylo zavřeno (Vorgangspassiv: *das Fenster wurde geschlossen*)
vs. *okno bylo zavřené* (Zustandspassiv: *das Fenster war geschlossen*).[27]

[27] http://www.etymos.de/sprachen/tschechisch/ ETYMOS

Des Weiteren verkompliziert im Hinblick auf Tschechisch im IR die schon erwähnte Fusion durch die Veränderung des Wortstamms die Flexion. Für das Wort *žena* (dt. „Frau“) können der Tab. 5 zehn verschiedene Wortformen entnommen werden:

*žena, žen**y**, žen**ě**, žen**u**, žen**ou**, že**n**, žen**ám**, žen**ách**, žen**ami**.*

Die gemeinsame Wurzel dieser Wortformen umfasst lediglich drei Buchstaben: *žen*. Betrachtet man das Verb *ženit* (dt. „verheiraten“), so fällt auf, dass wenn die für tschechische Verben typische Endung *–it* gestrichen wird, auch das Verb *ženit* (mit seinen ganzen Konjugationsformen) auf den gleichen Wortstamm wie *žena* reduziert wird.

Somit wäre die Verwendung der gemeinsamen Wurzel *žen* als Suchterm im IR ungeeignet. Der *Recall* würde ein enormes Ausmaß annehmen.

3.2 Fazit für das tschechische Information Retrieval

Zusammenfassend werden nun die Eigenschaften der tschechischen Sprache, die für das IR von Bedeutung sind und die daraus resultierenden Konsequenzen angeführt:

- Der tschechische Zeichensatz besteht aus lateinischen Buchstaben mit diakritischen Zeichen und muss während des ganzen IR-Prozesses unterstützt werden.
- Im Tschechischen kommen Komposita nur gering vor, sodass der aufwendige und komplizierte Prozess der Kompositazerlegung vermutlich vernachlässigt werden kann.
- Tschechisch hat eine starke Flexions- und Derivationsmorphologie. Diese sprachspezifische Besonderheit muss v.a. beim *automatischen Indexieren*, genauer gesagt beim *Stemmen*, beachtet werden. Der Prozedur des *Stemmens*, sowie dem *Stemmen* der tschechischen Sprache, wird im Kapitel 5.3 eine besondere Aufmerksamkeit geschenkt. Die reichen Wortvariationen führen zu einem hohen Speicheraufwand und durch die Fusions-Eigenschaft der Morpheme entsteht viel *Noise* und eine Sprengung des *Recalls*.

Kapitel 4

Forschungsinitiativen für informationslinguistische Ressourcen der tschechischen Sprache

Dieses Kapitel gibt einen Überblick über die verschiedenen Forschungsinitiativen, die sich mit informationslinguistischen Ressourcen für die tschechische Sprache befassen. Zunächst werden universitäre Projekte in der Tschechischen Republik vorgestellt und anschließend folgt ein kurzer Abriss, der die bedeutendsten Initiativen zusammenstellt, bei denen Tschechisch vertreten ist.

4.1 Universitäre Einrichtungen in der Tschechischen Republik

An der **Karls-Universität in Prag**[28] ist für den Forschungsbereich der Informationslinguistik das seit 1990 bestehende *Institute of Formal and Applied Linguistics (ÚFAL-Ústav formální a aplikované lingvistiky)*[29] zuständig. Die Forschungsarbeiten umfassen hauptsächlich Aktivitäten aus den Bereichen der *Maschinellen Übersetzung*, *Computerlinguistik*, *Korpora* und *Spracherkennung*. Zu den bedeutendsten und umfassendsten Projekten zählen:

- *Prague Dependency Treebank* (PDT) [30]
- *Czech Academic Corpus* (CAC) [31]
- *Multilingual Access to Large Spoken Archives* (MALACH) [32]
- *Valency Lexicon of Czech Verbs* (VALLEX 1.0)[33]

[28] http://www.cuni.cz/
[29] http://ufal.mff.cuni.cz/
[30] http://ufal.ms.mff.cuni.cz/pdt/index.html
[31] http://ufal.ms.mff.cuni.cz/REST/CAC/CAC.html
[32] http://www.clsp.jhu.edu/research/malach/
[33] http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0/

Eine mächtige informationslinguistische Ressource, die von der *Geisteswissenschaftlichen Fakultät* der Karls-Universität erstellt wurde, ist das *Tschechische Nationalkorpus* (TschNK)[34]. Dieses Korpus wurde im Rahmen dieser Arbeit für die Erstellung einer tschechischen Stoppwortliste verwendet und wird im Kapitel 5.1.1 dargestellt.

Die für informationslinguistische Forschungsaktivitäten zuständige Institution der **Masaryk-Universität in Brünn**[35] ist das *Natural Language Processing Laboratory*[36] der Fakultät für Informatik. Die Arbeiten liegen laut deren Aussage in den folgenden Bereichen:

1. *"The issues of synthesis and recognition of spoken language (Czech), dialogue systems. The results are further applied in the development of software tools for handicapped people, esp. visually impaired ones.*
2. *Lexical databases with relation to knowledge representation (Czech WordNet and Czech Lexical Database). Tools for viewing and editing dictionaries represented in XML format.*
3. *The problems of syntactic and semantic analysis based on Transparent Intensional Logic.*
4. *Development of syntactic parsers (partial and general) for Czech and their exploitation as disambiguators.*
5. *Improvement and further development of the tools for morphological analysis of Czech, solving the word derivation problems.*
6. *Building corpora, tagging and disambiguating corpus text, tools for corpus modification, maintenance, corpus managers and graphical interface for corpora.*
7. *Exploring techniques and methods of machine learning for the purpose of disambiguation of corpus data."*

Ein von der Masaryk-Universität mit der Westböhmischen Universität gemeinschaftlich organisiertes Projekt ist die jährlich stattfindende internationale Konferenz *Text, Speech and Dialogue* (TSD)[37]. Gegenstand dieser Konferenz ist in erster Linie die Verarbeitung der natürlichen Sprache, insbesondere Korpora, Texte, Sprachanalyse, Spracherkennung und deren Interaktion mit natürlichsprachigen Dialogsystemen. In diesem Jahr findet sie im Zeitraum vom 12. bis 16. September 2005 in Karlsbad statt.

[34] http://ucnk.ff.cuni.cz/english/
[35] http://www.muni.cz/
[36] http://nlp.fi.muni.cz/nlp/aisa/NlpEn/nlplab.html
[37] http://www.kiv.zcu.cz/events/tsd2005/

An der **Westböhmische Universität in Pilsen**[38] erforscht das *Institut für Kybernetik* [39] den Bereich der Informationslinguistik. Im Zentrum der Untersuchungen stehen die automatische Suche von Schlüsselwörtern, die natürliche Sprachverarbeitung und die Visualisierungsmöglichkeiten der tschechischen Sprache. Zu den Projekten an denen das Institut teilnimmt, zählen u.a. das bereits erwähnte MALACH-Projekt und das Projekt *The Czech Car Speech Corpus for TEMIC SDS, GmbH*.[40]

4.2 Europäische Forschungsinitiativen

Die hier angeführten Forschungsinitiativen haben gemeinsam, dass sie von der Europäischen Union gefördert werden und ihre Forschungsaktivitäten die tschechische Sprache umfassen.

MULTEXT East[41] (Multilingual Text Tools and Corpora for Central and Eastern European Languages) umfasst eine Serie von Projekten, die zum einen das Ziel haben, Standards und Spezifikationen für die Verarbeitung von linguistischen Korpora zu erarbeiten. Zum anderen verfolgen sie das Ziel, Tools, Korpora und linguistische Ressourcen, die diesen Standards entsprechen, zu entwickeln. Neben Tschechisch sind 18 weitere Sprachen Bestandteil dieses Projektes. Die Ergebnisse von MULTEXT sind frei verfügbar für den nicht-kommerziellen und nicht-militärischen Gebrauch.

Telri (Trans-European Language Resources Infrastructure) ist eine von der EU gegründete Initiative, die die Zusammenarbeit zwischen der Industrie und akademischen Forschungseinrichtungen fördern soll. Im Rahmen dieses Projektes werden Korpora, maschinenlesbare Wörterbücher und Lexika, sowie Software-Tools für die linguistische Datenverarbeitung erstellt.[42]

Balkan WordNet[43] hat zum Ziel semantische Relationen zwischen den Wörtern in jeder Sprache des Balkans darzustellen und diese in einem nächsten Schritt zu verbinden, um ein multilinguales semantisches Netz zu erstellen. Auf diese Weise entsteht für

[38] http://www.zcu.cz/index-en.html
[39] http://www.kky.zcu.cz/index.php?lang=en
[40] http://ui.zcu.cz/grants.php
[41] http://www.lpl.univ-aix.fr/projects/multext/
[42] http://www.telri.de
[43] http://www.ceid.upatras.gr/Balkanet/index.htm

jede Sprache ein sog. *WordNet.* Das WordNet wird für jede Sprache in einer Datenbank gespeichert und verbindet diese untereinander. Das Projekt *Balkan WordNet* ermöglicht u.a. somit den cross-lingualen Vergleich. Obwohl die Tschechische Republik aufgrund ihrer geographischen Situation nicht zu den Balkanländern gezählt wird, ist die tschechische Sprache Bestandteil der Untersuchungen im Rahmen dieses Projektes. Die weiteren untersuchten Sprachen sind Bulgarisch, Griechisch, Rumänisch, Türkisch und Serbisch.

ELSNET [44] (European Network of Excellence in Human Language Technologies) läuft seit April 1991 und wird an der University of Utrecht koordiniert. Dieses Projekt hat die Förderung der Entwicklung auf dem Gebiet der Sprachtechnologie zum Ziel, indem es für europäische Forschungseinrichtungen einen gemeinschaftlichen Pool an linguistischen Ressourcen für viele europäische und außereuropäische Sprachen bereitstellt.

ELAN[45] (European Language Activity Network) führt informationslinguistische Ressourcen (mit einem Umfang von mindestens je 1 Million Wörtern bzw. 5.000 Einträgen) für 31 (hauptsächlich) europäische Sprachen in einer gemeinsamen Oberfläche so zusammen, dass sie über identische Anfrageprozeduren gleichermaßen benutzt werden können.

[44] http://www.elsnet.org/
[45] http://solaris3.ids-mannheim.de/elan/

Kapitel 5

Ressourcen für die tschechische Sprache

Die meisten europäischen Sprachen (wie z.B. Französisch, Spanisch, Englisch, Deutsch oder Tschechisch) gehören zu dem indo-europäischen Sprachstamm und haben somit viele Eigenschaften gemeinsam, sodass Regelmäßigkeiten erkannt werden können. Beispielsweise werden Wortgrenzen (im Unterschied zu zahlreichen asiatischen Sprachen) durch Leerzeichen getrennt und verschiedene Wortvarianten eines Wortes entstehen durch das Anhängen von Suffixen an den Wortstamm.

Diese Regelmäßigkeiten können genutzt werden, um ein IR-System an eine bestimmte Sprache anzupassen. Hierzu werden erstrangig zwei informations-linguistische Ressourcen verwendet: zum einen eine allgemeine Stoppwortliste, zum anderen ein schnelles Stemming-Verfahren.

Die Stoppwortliste enthält nichtsignifikante Wörter, die von einem Dokument oder einer Anfrage vor Beginn des Indexierungsprozesses entfernt werden. Beim Stemming-Verfahren wird versucht Flexions- und Ableitungssuffixe zu entfernen, um die verschiedenen Wortvarianten eines Wortes auf den gleichen Wortstamm oder *Stem* zu reduzieren. Der Einsatz dieser Ressourcen im IR, sowie deren positiven Auswirkungen im IR-Prozess, werden in den Kapitel 5.2 und 5.3 erläutert.

Dabei sollte nicht vergessen werden, dass bei der Lösung dieses Problems für slawische Sprachen, insbesondere für die tschechische Sprache, durch die komplexere Morphologie im Vergleich zu Englisch (vgl. Kapitel 3.1), schwerer zu bewältigende Herausforderungen gestellt werden.

Dieses Kapitel präsentiert die für diese Arbeiten verwendeten und erstellten Ressourcen. Es ist folgendermaßen gegliedert: der erste Abschnitt 5.1 gibt zunächst eine kurze Einführung in die Thematik der Korpora als informations-linguistische Ressourcen. Dabei werden das „Tschechische Nationalkorpus“ und die in ihm enthaltenen Korpora, sowie der Korpus-Manager BONITO kurz vorgestellt.

Anschließend wird genauer auf das in dieser Arbeit verwendete Korpus SYN2000 und dessen Verwendung für die Erstellung von Stoppwortlisten eingegangen. Abschnitt 5.2 beleuchtet die Herangehensweise und die Verfahren für die Erstellung von allgemeinen Stoppwortwortlisten mit deren Hilfe im Rahmen dieser Arbeit eine Stoppwortliste für Tschechisch erstellt wurde. Anschließend erfolgen der Vergleich der vorhandenen Stoppwortlisten und deren Auswertung im Hinblick auf MIMOR. Im Abschnitt 5.3 werden der Universal-Stemmer *EgoThor* und die auf ihm basierende Stemming-Prozedur des polnischen Stemmers *STEMPEL* im Hinblick auf die Verwendung auf tschechische Texte vorgestellt und diskutiert. Abschnitt 5.4 führt als weiteres Ergebnis dieser Arbeit den intellektuell erstellten Text-Katalog für die tschechische Toplevel-Domain von WebCLEF an.

5.1 Korpora

Ein Korpus ist eine Menge von (geschriebenen oder gesprochenen) Texten, die meist in geschriebener Form und elektronisch gespeichert vorliegen und gegebenenfalls mit (Meta-) Kommentaren versehen und/oder linguistisch aufbereitet sind. (vgl. Walther von Hahn, 2004, 2).

John Sinclair (1991, 171) definiert ein Korpus folgendermaßen:

> "A corpus is a collection of naturally-occuring language text, chosen to characterize a state or variety of language."

In dieser Definition kommt ein entscheidender Aspekt hinzu: das gezielte Auswählen der Texte. Der Gedanke, dass ein Korpus zu einem bestimmten Zweck erstellt wird und einen bestimmten Ausschnitt einer Sprache darstellt, ist hervorzuheben. Wie Chomksy schon 1962 bemerkte, basieren alle Korpora auf bestimmte Auswahlkriterien:

> "Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite."[46]

[46] Chomsky 1962, zitiert nach McEnery (1996, S.8).

Chomskys Kritik ist auch heutzutage noch gerechtfertigt, da in einem Korpus niemals alle Variationen einer Sprache erfasst werden können. Um also mit einem Korpus einen möglichst großen Teil der in der jeweiligen Sprache vorkommenden Phänomene abzudecken, ist es notwendig, bei der Auswahl der Korpustexte ein breites Spektrum an Textformen- und Kategorien, welche im Idealfall auch quantitativ die Anzahl der in der Realität produzierten und rezipierten Texte widerspiegeln, zugrunde zu legen.

Ein Korpus wird als Ressource für empirische Sprachstudien in vielen Gebieten der Sprach- und Informationswissenschaften verwendet (Morphologie, Syntax, Semantik, Soziolinguistik, IR, etc.). Dabei stellen Korpus-Manager mächtige Tools für die Arbeit mit Korpora dar.

Laut Hahn (2004, 4) lassen sich folgende Korpustypen unterscheiden:

- *Schriftliche Textkorpora* liegen in geschriebener oder in transkribierter gesprochener Sprache vor und haben meistens als Grundeinheit Tokens.
- *Sprachsignalkorpora* enthalten Sprachsignale mit phonetischen und linguistischen Annotationen (Phonemgrenzen, Grundfrequenz, orthographische Transkription).
- *Multimodale Korpora* haben die gleichen Eigenschaften wie Sprachsignalkorpora, sind aber zusätzlich mit Mimik, Gestik und Bildern annotiert.
- *Strukturkorpora* (z.B. Baumbanken) beinhalten syntaktisch analysierte Sätze und haben als Grundeinheit den Satz.

Neben den Korpustypen gibt es noch technischen Aspekte, die die Korpora charakterisieren. Dazu gehören die Kodierung (in welchem Format das Korpus vorliegt), die Größe (gemessen als Anzahl der gespeicherten Zeichen, Tokens oder Sätze), die verschiedenen Möglichkeiten der Abfrage und die Art der enthaltenen Information, sowie die Zusammensetzung des Korpus (homogen, interessant für Terminografie und Lexikografie vs. heterogen, interessant für linguistische oder literarische Analyse), die dazugehörigen Metainformationen (Dokumentstruktur, Jahr, Autor, Tags, usw.) und die linguistischen Annotationen (Wortarten, Grundformen von Wörten, etc.).

5.1.1 Das Tschechische Nationalkorpus (TschNK)[47]

Seit 1994, dem Entstehungsjahr des TschNK, ist das Institut des Tschechischen Nationalkorpus („Ústav Českého národního korpusu“, ÚČNK) der Philosophischen Fakultät der Karlsuniversität Prag für dessen Pflege, Entwicklung und den damit verbundenen Aufgaben, wie bspw. die Arbeiten an der Korpuslinguistik, zuständig.

Das TschNK wurde in einer Projektarbeit für die weitere Erforschung der tschechischen Sprache erstellt und ist eine Sammlung von tschechischsprachigen Korpora, die Texte in elektronischer Form darbieten. Die Arbeit mit dem TschNK erfolgt online unter http://ucnk.ff.cuni.cz/, nach vorheriger Registrierung, für nicht kommerziellen Nutzen mit Hilfe der web-basierten Ressource BONITO. Der Programm-Manager BONITO wird im nächsten Kapitel vorgestellt.

Weiterhin besteht die Möglichkeit offline mit den Korpora des TschNK zu arbeiten. Dazu genügt die beim „ÚČNK“ käufliche CD-ROM[48]. Auf dieser CD-ROM befinden sich die zwei Korpora SYNEK (SYNchronoischer Elektronischer Korpus) und LITERA. SYNEK ist mit 10 Mio Wörtern die verkleinerte Ausgabe von SYN2000, wobei die Zusammenstellung von SYN2000 beibehalten wurde. Das Korpus LITERA enthält repräsentative Auszüge der tschechischen Prosa des 20. Jahrhunderts. Für die Arbeit mit den beiden Korpora eignet sich der Korpus-Manager BONITO.

Im TschNK sind verschiedene Korpora enthalten, die sich in ihrer Form (gesprochen, geschrieben, Dialekte/Mundart) und zeitlicher Herkunft ihrer Dokumente (zeitgenössisch, Alttschechisch) unterscheiden. Weitere Aspekte, die die Korpora unterscheiden, sind deren Domäne (Sparte bei Zeitungstext, Fachgebiet bei wissenschaftlichen Texten), das Alter der Texte, die Annotation und ihre Form.

[47] Český národní korpus (ČNK), http://ucnk.ff.cuni.cz/index.html
[48] Diese CD-ROM ist beim Institut für Angewandte Sprachwissenschaften der Universität Hildesheim hinterlegt.

Folgende Tabelle gibt eine Übersicht über die Korpora des TschNK:

<table>
<tr><th colspan="3">Das Tschechische Nationalkorpus (TschNK)</th></tr>
<tr><th colspan="2">Zeitgenössisch (synchronisch)</th><th>Alttschechisch (diachronisch)</th></tr>
<tr><td colspan="2">Die Sammlung zeitgenössischer tschechischer Texte (im SGML-Format) ist aufbereitet für die Suche mit dem Korpus-Manager.</td><td>Die Sammlung alttschechischer Texte ist untergliedert in: transkribierte Texte (2 Mio Wörter), transliterierte Texte (ca. 100 000 Wörter) und dialektale Texte (ca. 200 000 Wörter)</td></tr>
<tr><th>Gesprochene Korpora</th><th>Geschriebene Korpora</th><th>DIAKORP</th></tr>
<tr><td>▪ Prager Sprechkorpus
▪ Brünner Sprechkorpus</td><td>▪ SYN2000
▪ FSC2000
▪ PUBLIC
▪ SYNEK
▪ LITERA
▪ ORWELL</td><td>bietet eine Auswahl alttschechischer Texte von den ersten gefundenen Aufzeichnungen bis hin zu Texten der zeitgenössischen Korpora.</td></tr>
</table>

Tab.6 : *Übersicht über die Korpora des TschNK*

5.1.2 Der Korpus-Manager BONITO[49]

Um die Arbeit mit den Korpora zu erleichtern wurde eine spezielle Software, der sog. Korpus-Manager BONITO, von Pavel Rychlý für Suchanfragen entwickelt. Mit Hilfe von BONITO ist es möglich nach Wörtern und Wortkombinationen im Kontext zu suchen, sowie sich die Angabe der Termfrequenz im Korpus und im ursprünglichen Textdokument ausgeben zu lassen.

Des Weiteren ermöglicht BONITO weitere Operationen wie z.B. das alphabetische Sortieren und bei einigen Korpora sogar die Suche nach Wortform, Lemma, Tag und Wortgruppe. Lemmata oder Lexeme sind die Grundformen der lexikalischen Einheiten, unter denen man sie im Lexikon aufsuchen kann und werden im Lexikon der Sprachwissenschaft (1990) als:

„Eintrag, bzw. einzelnes Stichwort in einem Lexikon oder Wörterbuch“

definiert.

[49] BONITO ist eine web-basierte Ressource und kann unter http://ucnk.ff.cuni.cz/bonito heruntergeladen werden. Des Weiteren werden unter dieser URL ausführlich Ratschläge zum Umgang mit BONITO ausgeführt.

Lemmatisierung, im Sinne der Informationslinguistik, ist ein notwendiger Prozess zur Herstellung von Indices, Konkordanzen, Wortlisten, usw. für Textkorpora, da den einzelnen Wortformen eine einheitliche Leitform zugeordnet wird.

Ein Tag ist die morphologische Bezeichnung einer Wortform, zum Beispiel bei Substantiven: Genus, Kasus, Numerus. *Tagging* ist die verbreiteste Annotationsart von Korpora. In der Literatur werden hierfür auch die Termini *part-of-speech tagging* oder morphosyntaktische Annotation verwendet (vgl. McEnery 1996, S. 36).

5.1.3 Das Korpus SYN2000[50]

Das größte tschechische Korpus ist SYN2000. Es umfasst 100 Mio. Wörter (100000704 Tokens) und wurde im Oktober 2000 der Öffentlichkeit zugänglich gemacht. Seine Datenmenge erreicht nicht ganz die 2 GB-Grenze. SYN2000 setzt sich zusammen aus 3303 Texteinheiten, 7 639 ganzen Sätzen und 1 763 818 Types. Dieses Korpus besteht aus ganzen Texten, die für speziell für Forschungszwecke auf dem Gebiet der geschriebenen Sprache aufbereitet wurden. Was die linguistischen Annotationen betrifft, so ist SYN2000 lemmatisiert und morphologisch gekennzeichnet (getaggt).

SYN2000 ist ein zeitgenössisches Korpus, dies bedeutet, dass es in dem Tschechisch verfasst ist, wie es heutzutage gesprochen, bzw. geschrieben wird. Hauptsächlich sind in ihm Texte aus den Jahren 1990 bis 1999 zu finden.

Im Korpus sind auch bedeutende Werke der tschechischen Literatur enthalten, die vor 1990 entstanden sind (z.B. „Krakatit“ von Karel Čapek, oder „Zbabělci“ von Josef Škvoreck). Für ältere Texte galt die Bedingung für die Aufnahme in das Korpus, dass der Autor vor 1880 geboren sein musste.

[50] *Český národní korpus*: http://ucnk.ff.cuni.cz

Folgende Graphik veranschaulicht die vor allem aus publizistischen Texten bestehende Zusammenstellung des Korpus:

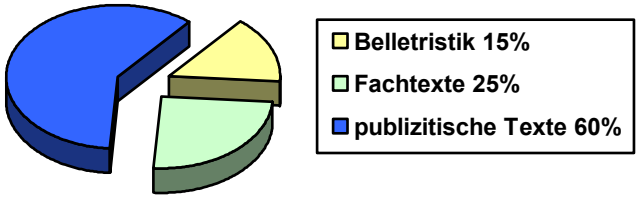

Abb. 10: *Zusammenstellung des Korpus SYN2000*

Weiteren Aufschluss über die Texte gibt das nächste Schaubild, das die Herkunft der Fachtexte illustriert.

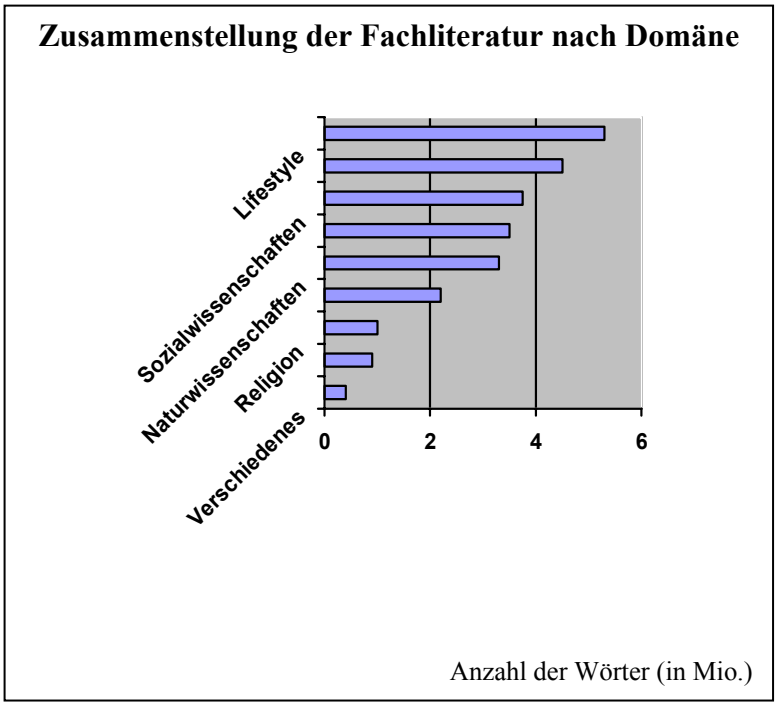

Abb. 11: *Zusammenstellung der Fachliteratur des TschNK nach Domäne*

Im nächsten Kapitel wird beschrieben wie die allgemeinen Stoppwortlisten für die tschechische Sprache mit Hilfe des Korpus SYN2000 erstellt wurden. Folgende Kriterien sprachen für die Verwendung von SYN2000: Die allgemeinen Stoppwortlisten sollten die am häufigsten verwendeten Wörter in der tschechischen Sprache enthalten. Voraussetzung hierfür ist, dass eine für die tschechische Sprache

möglichst repräsentative Textsammlung auf Termfrequenzen untersucht wird. Das Kriterium der Repräsentativität ist bei SYN2000 erfüllt, da es sich um eine sehr große Kollektion handelt und die darin enthaltenen Texte nicht domänenspezifisch, sondern allgemein gehalten sind.

Maßgebend war auch die Tatsache, dass es sich um ein geschriebenes, zeitgenössisches Korpus handelt, dessen vorwiegende Textsorte publizistische Texte sind, da die im Rahmen dieser Arbeit erstellten Stoppwortlisten ihren späteren Einsatz in MIMOR@CLEF finden werden und somit für die Entfernung von Stoppwörtern in der dort am stärksten vertretenen Textsorte (Zeitungsartikeln) verwendet werden.

5.2 Stoppwortlisten

Es gibt Wörter, die sehr häufig auftreten und folglich für den Dokumenteninhalt keine oder eine nur sehr geringe Bedeutung haben. Diese Wörter werden im IR als *Stoppwörter* bezeichnet. Sie werden bei einer Volltextindexierung nicht beachtet, da sie für gewöhnlich keine Relevanz für die Erfassung des Dokumentinhalts besitzen.

Beispiele für solche Stoppwörter sind:

- Artikel
- Konjunktionen
- Präpositionen
- Fragewörter
- Modalverben
- und Negationen (in deutschsprachigen Dokumenten z.B.: „nicht“). [51]

Das häufige Auftreten der Terme (auch „hohe Frequenz“) bedeutet, dass sie innerhalb eines Dokumentes und in der gesamten Dokumentenkollektion sehr zahlreich auftreten. *Stoppwörter* würden aus diesem Grund bei der Erschließung der Dokumente einen hohen Aufwand verursachen. Des Weiteren übernehmen *Stoppwörter* vor allem

[51] Abhängig von den zu erschließenden Dokumenten können Stoppwörter, bzw. Stoppwortlisten auch mehrsprachig vorliegen (vgl. die multilinguale Stoppwortliste von Jensen (2005)).

grammatikalische, bzw. syntaktische Funktionen und lassen daher keine Rückschlüsse auf den Inhalt des Dokumentes zu.

Es macht keinen Sinn, eine Suchanfrage mit einem *Stoppwort* zu stellen, da hierfür fast alle Dokumente relevant sind. Die Ergebnismenge würde nahezu jedes Dokument des Bestandes enthalten. Ein solches Suchergebnis wäre für den Anwender nutzlos. Moderne Suchmaschinen filtern *Stoppwörter* heraus. So werden allgemeine Wörter und Buchstaben, die als *Stoppwörter* bekannt sind, ignoriert. Dies bedeutet, dass Wörter wie *"aber"* und *"der"*, wie auch bestimmte einzelne Zahlen und Buchstaben automatisch übergangen werden, weil diese Begriffe nur wenig bei der Einschränkung der Suche helfen und gleichzeitig die Suchgeschwindigkeit bedeutend verlangsamen. Durch die Eliminierung der Stoppwörter im Index wird die Größe der invertierten Liste verringert. Laut Jacques Savoy (1999, 945) kann durch die Verwendung einer *Stoppwortliste* mit einer Reduzierung von 30 bis 50% gerechnet werden. Somit dienen *Stoppwörter* der Steigerung der Effizienz des IR-Prozesses.

Die Länge einer *Stoppwortliste* im Einsatz bei IR-Systemen variiert stark. So enthält z.B. die von zahlreichen CLEF-Teilnehmern benutzte deutsche *Stoppwortliste*, die zum Download auf der Homepage der *Université de Neuchâtel* bereit steht, 603 Wörter.[52] Wohingegen die im *German Analyzer* von Lucene enthaltene deutsche Default-*Stoppwortliste* nur aus 48 Wörtern besteht.[53] Die Länge, sowie die Zusammensetzung der Liste sind jedoch domänenspezifisch, da bspw. im juristischen Kontext die Präpositionen und Zahlwörter eine entscheidende Rolle spielen und in einem Wirtschaftsarchiv andere Wörter häufiger auftreten als in einer medizinischen Datenbank.

Ziel dieser Arbeit ist unter anderem die Generierung einer erschöpfenden Stoppwortliste für Tschechisch, damit diese in die Erweiterung MIMORs für Tschechisch eingebunden werden kann. Die Qualität der Stoppwortliste ist laut Savoy (1999, 945) maßgeblich entscheidend für eine erfolgreiche Indexierung. Das endgültige Ziel wird erreicht, wenn angemessene Indexterme für die Dokumente und die Anfrage vergeben werden können, sodass die gewünschte Retrievalqualität erreicht werden kann.

[52]http://www.unine.ch/info/clef/
[53]http://svn.apache.org/repos/asf/lucene/java/trunk/contrib/analyzers/src/java/org/apache/lucene/analysis/de/GermanAnalyzer.java

5.2.1 Richtlinien für Vorgehensweise

Als Richtlinien für die Erstellung der tschechischen Stoppwortlisten dienten die Arbeiten von Jacques Savoy[54], die wiederum auf denen von Christopher J. Fox[55] beruhten. Laut deren Vorgehensweise sollten im ersten Schritt die Terme, die im Korpus vorkamen, nach ihrer Häufigkeit („Termfrequenz“) sortiert werden, wobei die 200 häufigsten Terme, also die mit der höchsten Termfrequenz, für die Liste ausgewählt werden sollten.

Im zweiten Schritt wurden von der Liste alle Zahlen (z.B. „1998“, „1“), sowie alle Substantive und dazugehörigen Adjektive, die im starken Bezug zu der Hauptthematik der Kollektion standen, entfernt. Beispielsweise wurde in der von Jacques Savoy angeführten Studie zum AMARYLLIS-Projekt u.a. der Korpus OFIL verwendet, der ausgewählte Artikel aus der französischen Tageszeitung *Le Monde* enthielt. Die Wörter „France“ und „Président“ gehörten zu den 200 Termen, die aufgrund ihrer hohen Frequenz aus diesem Korpus extrahiert wurden und fanden sich auf der 66. bzw. 69. Stelle der Liste wider. Diese Wörter hatten einen starken thematischen Bezug zu der Zeitung *Le Monde* und wurden von der Liste gelöscht, da angenommen wurde, dass solche Wörter nur unter bestimmten Umständen als Indexterme nützlich sein könnten.

Im dritten und letzten Schritt wurden Wörter der Stoppwortliste hinzugefügt, die aus informationswissenschaftlicher Sicht keinen Informationsinhalt trugen. Dies geschah auch, wenn diese nicht zu den 200 häufigsten Wörtern des Korpus zählten. Solche Wörter waren z.B. Personal- und Possessivpronomen und Konjunktionen. Die resultierende Stoppwortliste enthielt somit eine große Anzahl an Pronomina, Artikeln, Präpositionen und Konjunktionen.

[54]Savoy, Jacques (1999), *A Stemming Procedure and Stopword List for General French Corpora*, In: Journal of the American Society for Information Science 50(10):944–952.
[55]Fox, C. (1990), *A stop list for general text.* ACM-SIGIR Forum, 24, 19–35.

5.2.2 Ausgangsbasis und eigene Vorgehensweise

Die Vorgehensweise von Savoy (1999) und Fox (1990) wurde in dieser Arbeit nur teilweise übernommen, da andere Ausgangsbedingungen herrschten. Ausschlaggebend war in erster Linie, dass bereits vier Stoppwortlisten für Tschechisch vorlagen und als Ausgangsbasis fungierten. Für jede Stoppwortliste wurden Häufigkeitstabellen mit den folgenden Angaben erstellt: *absolute Termfrequenz*, *relative Termfrequenz*, kumulierte *relative Termfrequenz*, *Prozenthäufigkeit* und *kumulierte Prozenthäufigkeit*.

In den folgenden Abschnitten werden die jeweiligen Stoppwortlisten vorgestellt und die Erarbeitung der Häufigkeitstabellen am Beispiel der Stoppwortliste aus dem *Tschechischen Analyzer* beschrieben.

Die erste Stoppwortliste stammt aus dem *Tschechischen Analyzer* für Lucene[56] und enthält 172 Terme. Erstellt wurde der Analyzer von Lukáš Zapletal.[57] Bei der im *Tschechischen Analyze*r integrierten Stoppwortliste handelt es sich um die Default-Stoppwortliste[58], die im Programm-Code enthalten ist und verwendet wird, falls keine externe Stoppwortliste angegeben wird.

Die zweite Stoppwortliste wurde von Vašek Nemčík mit dem Korpus DESAM[59] erstellt, das aus 251 805 Tokens besteht. DESAM ist ein Projekt der Fakultät für Informatik an der Masaryk Universität in Brünn, dessen Suchmaske unter http://www.fi.muni.cz/~pary/korp/desq.cgi aufgerufen werden kann. Die Stoppwortliste enthält 75 Terme, sowie die Werte der dazugehörigen relativen Prozenthäufigkeit.

Die dritte Stoppwortliste wurde ebenfalls von Vašek Nemčík erstellt[60]. Für deren Anfertigung hat er aber mit zwei Korpora gearbeitet: ESO (53 389 437 Tokens) und dem TschNK. Mit dem Korpus ESO kann ebenfalls online gearbeitet werden: http://www.fi.muni.cz/~pary/korp/query.cgi. Diese Stoppwortliste besteht aus 76 Termen und gibt Aufschluss über deren relative Prozenthäufigkeit, sowie Angaben

[56] Lucene ist ein Open Source-Projekt der *Apache Software Foundation* (ASF)
[57] weitere Informationen zu dem Autor unter: http://lukas.zapletalovi.com/
[58] http://svn.apache.org/repos/asf/lucene/java/trunk/contrib/analyzers/src/java/org/apache/lucene/analysis/cz/CzechAnalyzer.java
[59] persönliche Kommunikation, (vgl. Nemčík 2005)
[60] persönliche Kommunikation, (vgl. Nemčík 2005)

über das Vorkommen am Anfang eines Satzes oder innerhalb eines Satzes des entsprechendes Termes.

Die vierte vorliegende Stoppwortliste entstand im Rahmen der Magisterarbeit von O. Artemenko & M. Shramko (2005), *Entwicklung eines Werkzeugs zur Sprachidentifikation in mono- und multilingualen Texten*. Sie enthält 247 Terme und die Werte der absoluten, der relativen, der inversen und der kumulierten relativen Termfrequenz. Als Ausgangsbasis für die Ermittlung der Häufigkeiten diente eine 200kb große Textdatei, die Zeitungsartikel aus der tschechischen Tageszeitung *Lidové noviný* enthielt.

Im Unterschied zu den drei anderen Stoppwortlisten musste bei der Liste des *Tschechischen Analyzers* die Codierung des Zeichensatzes geändert werden. Die Stoppwörter waren im Unicode-System (in hexadezimaler Angabe) codiert, in Anführungsstrichen gesetzt und durch Kommata getrennt.[61] Der für Tschechisch notwendige Zeichensatz ist „Latin Extended-A" und umfasst den Zeichenbereich von „U+0100" bis „U+017F". Zum Beispiel steht „\u0159" für den tschechischen Buchstaben „ř".

Im den folgenden Abschnitten werden anhand der Default-Stoppliste des *Tschechischen Analyzer*s die weiteren Verfahrensweisen exemplarisch erklärt. Die drei weiteren Stoppwortlisten wurden der gleichen Prozedur unterzogen.

In einem ersten Schritt wurde die Stoppwortliste im Text-Editor Bloc-notes in Form gebracht, sodass pro Zeile ein Wort stand. Diese Liste wurde in Excel übertragen und die Sonderzeichen, die im Unicode erschienen, wurden ersetzt.

Daraufhin wurde diese Liste alphabetisch sortiert, um die anschließende Suche im TschNK zu erleichtern. Mit Hilfe des Programm-Managers BONITO wurde die *absolute Termfrequenz* (auch „*einfache Termfrequenz*" oder „*Termhäufigkei*t") für jedes Wort im TschNK ermittelt. Im Falle der Stoppwortlisten von Nemčík und Artemenko, Shramko (2005) waren zwar die Häufigkeitsverteilungen mitangeführt,

[61] Unicode ermöglicht die Codierung von Zeichen oder Elementen aller bekannten Schriftkulturen und Zeichensysteme (vgl. http://www.unicode.org/charts/PDF/U0100.pdf)

diese stammten aber nicht aus SYN2000 und konnten somit nicht für einen Vergleich mit den anderen Stoppwortlisten herangezogen werden. Die Termfrequenzermittlung im TschNK wurde durchgeführt, um die in den Listen angeführten Stoppwörter auf „Stoppwortwürdigkeit“ zu testen. Dabei ist zu beachten, dass, um die tatsächliche Termfrequenz eines Wortes als Ergebnis zurückzubekommen, es notwendig ist, sich alle Schreibweisen eines Wortes (nur Groß- und Kleinschreibung betreffend, weitere Rechtschreibfehler wurden sinnvollerweise nicht berücksichtigt) von BONITO ausgeben zu lassen und deren *absolute Termfrequenz* im nächsten Schritt miteinander zu verrechnen ist.

Denn bei der Eingabe „ale“ erhält man die Antwort, dass dieser Term 326187 Mal im Korpus auftritt. Darin nicht enthalten sind die Schreibvarianten „Ale“ (tritt am Anfang des Satzes auf, Termfrequenz: 62476) und „ALE“ (in Grossbuchstaben, Termfrequenz: 348). Somit müssen die Termfrequenzen der Schreibvarianten addiert werden, um für einen Term die *absolute Termfrequenz* zu erhalten. Im Beispiel ergibt sich für den Term „ale“, wenn seine dazugehörigen Schreibvarianten berücksichtigt werden, folgende *absolute Termfrequenz*: 389011 (326187+62476+348= 389011).

Verschiedene Schreibvarianten eines Terms	***absolute Termfrequenz* pro Schreibvariante**	**standartisierte Schreibweise des Terms**	**Summe der *absoluten Termfrequenz***
a	2690157	**a**	2855855
A	165698		
aby	163991	**aby**	170612
Aby	6570		
ABY	51		
aj	3828	**aj**	4118
Aj	215		
AJ	75		
ale	326187	**ale**	389011
Ale	62476		
ALE	348		

Tab. 7: *Beispielhafter Auszug der Häufigkeitstabelle für die Berechnung der absoluten Termfrequenz*

Man könnte davon ausgehen, dass in der ersten Schreibvariante alle weiteren Varianten enthalten sind. Dass diese Annahme aber falsch ist, kann durch den

Gegenbeweis belegt werden, dass in einigen Fällen, die zweite (bei der nur der erste Buchstabe großgeschrieben wird) oder die dritte (in Grossbuchstaben geschriebene) Schreibvariante eine höhere Termfrequenz hat als die erste. Ein Beispiel hierfür ist das Wort „atp“, was die Abkürzung für „a tak podobně“, was soviel bedeutet wie „und ähnlich(es)“ (vgl. Tab. 8). Obwohl diese Abkürzung normalerweise in Kleinbuchstaben geschrieben wird (vgl. Velký Kapesní Německo-Česky, Česko-Německy Slovník 2000, 367), tritt im SYN2000 Korpus die Schreibweise „ATP“ fast doppelt so häufig auf, als die kleingeschriebene.

Verschiedene Schreibvarianten eines Terms	***absolute Termfrequenz* pro Schreibvariante**	**standartisierte Schreibweise des Terms**	**Summe der *absoluten Termfrequenz***
atp	651	**atp**	1995
Atp	73		
ATP	1271		

Tab.8: *Beispielhafter Auszug der Häufigkeitstabelle für eine mögliche falsche Berechnung der absoluten Termfrequenz durch verschiedene Schreibvarianten*

Eine weitere mögliche Fehlerquelle, die die Berechnung der Termfrequenz verfälschen kann, tritt auf, wenn eine nicht ganz korrekte, aber im Korpus existierende Schreibvariante (nur Groß- und Kleinschreibung betreffend) eines Wortes verwendet wird. Hier im Beispiel ist es die eigentlich unzulässige Form „PaK“.

Verschiedene Schreibvarianten eines Terms	***absolute Termfrequenz* pro Schreibvariante**	**standartisierte Schreibweise des Terms**	**Summe der *absoluten Termfrequenz***
pak	98675	**pak**	118559
Pak	19845		
PAK	39		
PaK	5		

Tab. 9: *Beispielhafter Auszug der Häufigkeitstabelle für die möglich falsche Berechnung der absoluten Termfrequenz durch außergewöhnliche Schreibvarianten*

In einem weiteren Schritt wurde mittels der *absoluten Termfrequenz* (hier der Wert, der sich als *Summe der absoluten Termfrequenzen* ergibt) und der *Gesamtzahl der*

Terme im Korpus die *relative Termfrequenz* für jeden Term berechnet. Für die *relative Termfrequenz* gilt folgende Formel:

$$relative\,Termfrequenz = \frac{absolute\,Termfrequenz}{Gesamtzahl\,der\,Terme\,in\,der\,Dokumentenmenge}$$

Für das Beispiel „ale“ wird die *relative Termfrequenz* folgendermaßen berechnet:

$$relative\,Termfrequenz = \frac{Summe\,der\,absoluten\,Termfrequenz}{Gesamtzahl\,der\,Terme\,im\,Korpus} = \frac{389011}{100000000} = 0{,}00389011$$

Daraus ergibt sich folgender Eintrag in der Ergebnistabelle:

Verschiedene Schreibvarianten eines Terms	***absolute Termfrequenz* pro Schreibvariante**	**standartisierte Schreibweise des Terms**	**Summe der *einfachen Termfrequenz***	***relative Termfrequenz***
ale	326187	**ale**	389011	0,00389011
Ale	62476			
ALE	348			

Tab. 10: *Auszug der Häufigkeitstabelle für die Berechnung der relativen Termfrequenz*

Die Häufigkeitstabellen ermöglichen die Berechnung der Abdeckung eines Korpus durch die Stoppwortliste. An dieser Stelle wird deutlich, weshalb die verschiedenen Schreibweisen eines Terms berücksichtigt werden müssen, um die absolute Termfrequenz zu berechnen. Mit Hilfe der bisher errechneten Häufigkeiten werden nun drei weitere Häufigkeiten berechnet, um die Abdeckung zu ermitteln.

Dazu wird die bis jetzt noch alphabetisch sortierte Liste nach der *absoluten*, bzw. *relativen Termfrequenz* sortiert, um als Ergebnis eine gerankte Liste zu erhalten. Die zwei ersten Spalten (**Verschiedene Schreibvarianten eines Terms** und ***absolute Termfrequenz* pro Schreibvariante**) sind in dieser Liste für die Auswertung nicht mehr von Interesse und werden aus diesem Grund gelöscht. Stattdessen werden drei weitere Maße für die Berechnung von Frequenzen hinzugefügt:

- die *kumulierte relative Häufigkeit:* $f_{cum}(k) = \sum_{k=1}^{m} f_k$

  Die *kumulierte Häufigkeit* gibt die einfache Aufsummierung der jeweiligen Häufigkeiten - in diesem Falle der relativen Häufigkeiten - wieder.

- die *Prozenthäufigkeit:* $\%_k = \frac{f_k}{n} \times 100\%$

  Die *Prozenthäufigkeit* drückt hier den prozentuellen Anteil der absoluten Häufigkeit aus.

- und die *kumulierte Prozenthäufigkeit*. $f_{cum\%}(k) = \sum_{k=1}^{m} \%_k$

  Die *kumulierte Prozenthäufigkeit* ist die Addition der vorangegangen Prozenthäufigkeiten.[62]

5.2.3 Auswertung der Ergebnisse im Hinblick auf MIMOR@CLEF

Die oben eingeführten Häufigkeiten werden in die Häufigkeitstabellen der jeweiligen Stoppwortlisten übernommen, sodass diese im nächsten Schritt verglichen werden können.

Als erstes wurden die Terme der jeweiligen Stoppwortlisten miteinander verglichen und in einer neuen Stoppwortliste zusammengestellt. Diese Stoppwortliste enthielt die Terme aller benutzten Stoppwortlisten (198 Terme) und hatte eine kumulierte Prozenthäufigkeit von 29,369784 %. Dies bedeutet, dass die in ihr enthaltenen Stoppwörter fast 30% des SYN2000 im TschNK abdeckten. Diese zusammengestellte Stoppwortliste war die neue Ausgangsbasis für weitere Operationen.

[62] http://www.kfunigraz.ac.at/hgrwww/script.old/text732.html

Die nächsten Schritte orientierten sich an der in Kapitel 5.2.1 beschriebenen Vorgehensweise von Savoy und Fox. Es wurde überprüft, ob in der Liste Zahlen vorkamen. Bis auf zwei Ausnahmen war dies nicht der Fall. Die erste Ausnahme ist „jeden", die ausgeschriebene Form von „1". „Jeden" kann zum einen die Zahl „1" bedeuten, zum anderen ist „jeden" auch der männliche unbestimmte Artikel im Singular. Aus diesem Grund wurde dieser Term nicht gestrichen. Der Term „dva" dagegen wurde gestrichen, da er nur die Bedeutung der Zahl „2" hat.

Savoy und Fox haben bei ihrer Vorgehensweise Terme, die in einem starken Bezug zur Hauptthematik der Kollektion standen, gelöscht. In dieser Stoppwortliste wurden solche Terme jedoch lediglich durch Fettschreibung gekennzeichnet. Dies geschah aus dem Grund, dass solche Wörter unter bestimmten Umständen doch als Indexterme nützlich sein können. Der letztendliche Nutzen der Indexterme bedarf weiterer Tests und kann nur ermittelt werden, wenn die Stoppwortliste im IR-System zum Einsatz kommt. Genauer gesagt, wäre in zukünftigen Tests zu testen, inwieweit sich das Ergebnis verändert, wenn die gekennzeichneten Terme weggelassen werden.

Im nächsten Schritt wurde eine Liste aus Personal- und Possessivpronomen, Präpositionen und Konjunktionen für die tschechische Sprache erstellt. Dabei wurden Terme übernommen, die noch nicht in der Stoppwortliste enthalten waren. Für die Terme dieser Liste wurde im TschNK die absolute Frequenz ermittelt.[63] Diese Liste enthält 40 Terme (in standardisierter Form) und wurde der eigenen, zusammengestellten Stoppwortliste hinzugefügt.

Die resultierende Stoppwortliste für die tschechische Sprache enthält 239 Terme und deckt mit ihrer kumulierten Prozenthäufigkeit 30,49% des SYN2000 im TschNK ab.

[63] Auch hier mussten die verschiedenen Schreibvarianten beachtet werden.

Folgende Graphik veranschaulicht den Verlauf der Abdeckung:

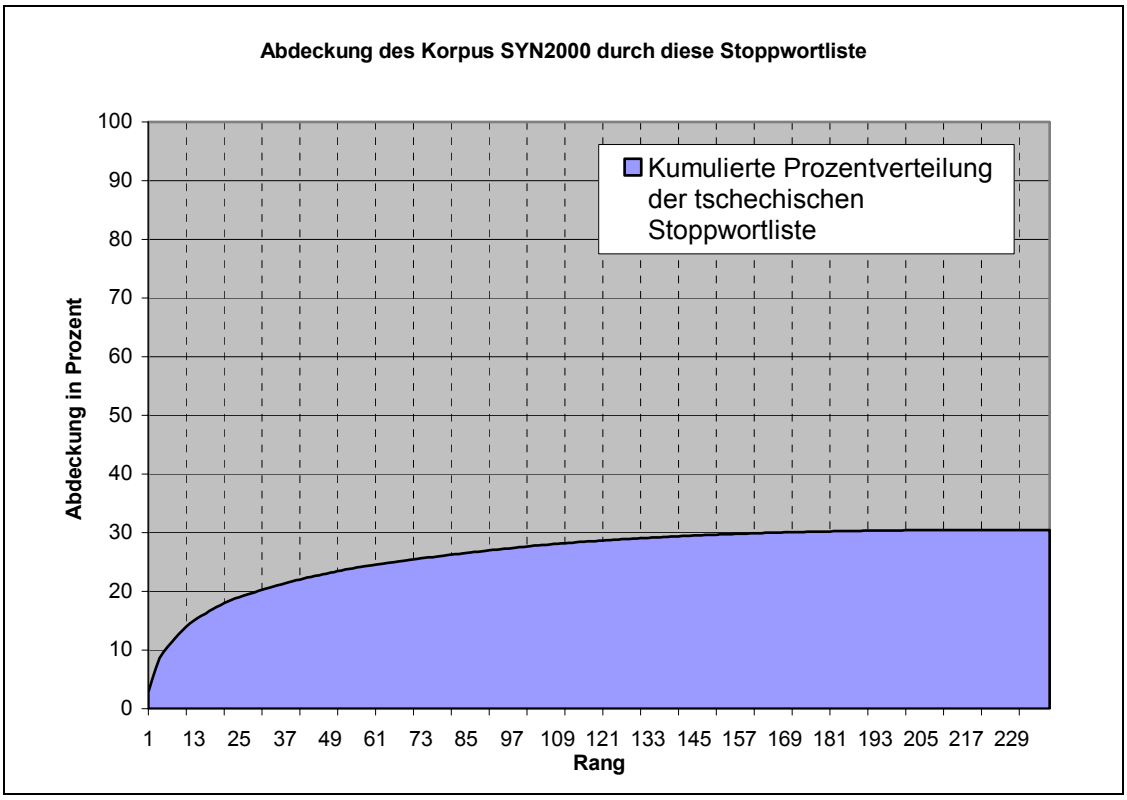

Abb.12 : *Verlauf der Abdeckung des Korpus SYN2000 durch die tschechische Stoppwortliste*

Die ersten fünf Terme decken bereits 9,65% des Korpus ab, die ersten zehn Terme 13,36%. Ab dem 130. Term steigt die Abdeckung nur noch geringfügig an.

Werden die Stoppwörter nach ihrer Häufigkeit sortiert, so wird auch das *Zipfsche Gesetz* bestätigt. Dieses empirisch oft beobachtete Gesetz wird häufig als Referenz bemüht, wenn auf die Tatsache, dass viele Wörter selten und wenige Wörter häufig vorkommen, Bezug genommen wird. Es besagt also u.a., dass die häufigsten Wörter einen großen Prozentsatz des Korpus abdecken.[64] Somit sind die häufigsten Wörter in einer Sprache von geringer Aussagekraft und werden deshalb in einer Stoppwortliste erfasst, um nicht als Indexterme eingesetzt zu werden.

Das *Zipfsche Gesetz* beschreibt den Zusammenhang zwischen der Vorkommenshäufigkeit eines Wortes und dessen Rangplatz[65] in einer nach dieser

[64] http://www.ais.fraunhofer.de/~leopold/Zipfkurz.pdf
[65] Der Rangplatz ergibt sich durch die Vorkommenshäufigkeit (hier: absolute Termfrequnez).

Häufigkeit sortierten Wortliste. Dabei ist das Produkt aus Rangplatz und dem Wert der absoluten Häufigkeit eine Konstante. Somit ist die Häufigkeit eines Wortes proportional zu seiner Rangstelle.[66] Abb.13 zeigt die Verteilungskurve der 80 ersten Terme in der tschechischen Stoppwortliste (durch die Prozenthäufigkeit ausgedrückt) und den Verlauf der Zipfschen Verteilung.[67]

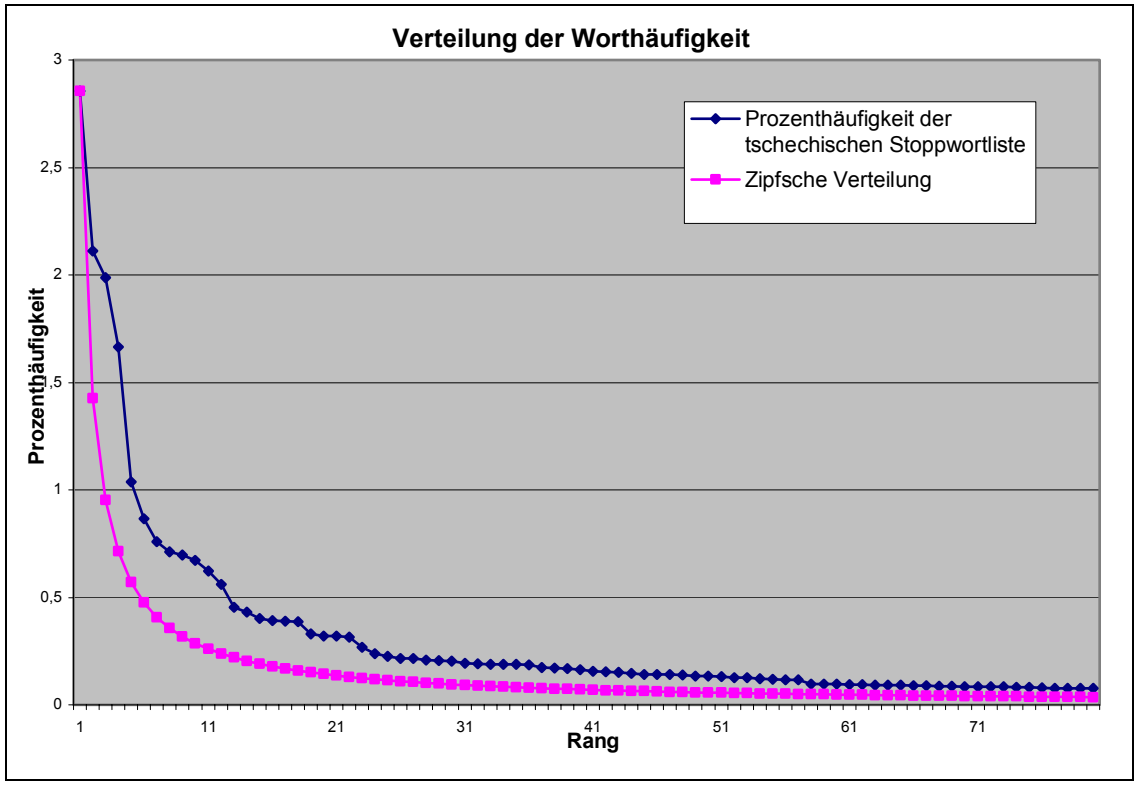

Abb. 13 : *Verteilung der Worthäufigkeit*

Wie jedes empirische Gesetz ist auch das Zipfsche Gesetz nur näherungsweise gültig. So illustriert die Graphik, dass die Verteilung der Worthäufigkeiten annähernd der Zipfschen Verteilung gehorcht. Weiterhin ist bei der Betrachtung der Kurven die Tatsache hervorzuheben, dass sich die Verteilungskurve der erstellten tschechischen Stoppwortliste für alle Werte oberhalb der Zipfschen Verteilungskurve liegt. Somit erzielen die in dieser Arbeit erfassten Stoppwörter eine bessere Abdeckung des Korpus SYN2000 als die Wörter der empirisch berechneten Verteilungskurve von Zipf.

[66] Ferber, Reginald (2003) *Information Retrieval. Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web.*dpunkt.verlag: Heidelberg.
http://information-retrieval.de/irb/ir.part_1.chapter_3.section_6.topic_3.subdiv1_1.html#Satz_1
[67] Es wäre nicht sinnvoll gewesen, mehr als 80 Terme durch die Verteilungskurve darzustellen, da ab ca. dem 70. Term sich der Kurvenverlauf nur noch minimal ändert, indem er asymptotisch gegen Null geht.

5.3 Stemmer

Ein *Stemmer* ist ein Programm oder ein Algorithmus, der die verschiedenen morphologischen Formen eines Terms auf einem gemeinsamen Kern, den sog. S*tem* reduziert. Dieser *Stem* muss nicht unbedingt ein in der Sprache existentes Morphem oder Lexem sein.

Das Konzept des *Stemmings* wurde in IR-Systemen seit ihren Anfängen angewandt (vgl. Gerald Kowalski 1997, 67) und bezeichnet das sprachspezifische Verfahren, mit dem verschiedene morphologische Varianten eines Wortes auf ihren gemeinsamen Wortkern zurückgeführt werden. Verschiedene Varianten eines Wortes können laut Frakes (1992, 161ff) entstanden sein durch:

- Komposition,
- Dekomposition[68],
- Flexion,
- und durch das Hinzufügen von Affixen (Präfix, Suffix, Infix).

Der *Stemmer* entfernt Flexionsendungen und Derivationssuffixe[69]. Er ermöglicht zugleich die Verknüpfung von ähnlichen Index- und Suchtermen, da die verschiedenen Wortformen auf den gleichen *Stem* reduziert wurden.

Ein englischer *Stemmer* sollte zum Beispiel den String “stemmer” (oder auch "stemming", "stemmed" etc.) als abgeleitete Form von "stem" erkennen. Englische Stemmer sind ziemlich trivial gebaut, da die englische Sprache nicht stark flektierend ist und kaum über Komposita verfügt. Der Term "dries" wird als dritte Person Singular Präsens des Verbes "dry" identifiziert und "axes" als Pluralform von "ax" (genauso wie "axis").

[68] laut Lexikon der Sprachwissenschaft : „Bezeichnung für mehr als zweigliedrige Komposita“, z.B.: Hundehalsband und Busfahrkartenautomat

[69] Bei einigen *Stemming-Verfahren* werden weitere „Endungsformen“ abgetrennt, wie z.B. bei statistischen Verfahren, die pauschal die letzten Buchstaben abtrennen.

Stemmer werden komplizierter, sobald die Morphologie, die Orthographie und der Zeichensatz der Zielsprache komplexer werden. So ist zum Beispiel ein tschechischer Stemmer komplexer als ein englischer, weil es im Tschechischen z.B. mehr mögliche Deklinationen für Substantive gibt.

Das *Stemming* ist von der *Lemmatisierung* zu unterscheiden, denn im Unterschied zum Stemming, wird bei der *Lemmatisierung* versucht, den Term auf seine "Lexikon-Form", *Lemma*, zu reduzieren. Es werden dabei alle Flexionsendungen entfernt und Umlaute, sowie graphemische Veränderungen werden in eine Standardform umgeformt. Der resultierende Output existiert als selbständiges Wort in der Sprache. *Lemmata* haben eine besonders hohe Bedeutung in stark flektierenden Sprachen zu denen auch das Tschechische zählt (vgl. Kapitel 3.1). Folgende Beispiele veranschaulichen das Verfahren der *Lemmatisierung* im Deutschen:

- Der Begriff „Freiheiten“ wird auf die Singularform „Freiheit“ reduziert, wobei die Pluralendung „-en“ entfernt wird.
- Beim Begriff „Umsätze“ kommt hinzu, dass der für die Pluralform typische Umlaut auf ein „a“ reduziert wird, sodass sich der Begriff „Umsatz“ ergibt.

Stemming wird im IR eingesetzt, da es zahlreiche Vorzüge bietet. In den Anfängen war das ursprüngliche Ziel des *Stemmings* die Leistung[70] eines IRS zu verbessern und weniger Systemressourcen zu benötigen, indem die Zahl der Wörter, die im System enthalten waren, reduziert wurden. Laut Frakes (1992, 131) kann eine Reduzierung des Speicherbedarfs für das Vokabular um bis zu 50% erfolgen.

Mit der ständig wachsenden Speicherkapazität und Leistungsfähigkeit der Computer hat aber die Bedeutung des *Stemmings* für die Performanz mit der Zeit abgenommen. *Stemming* Algorithmen werden nun vorrangig wegen der möglichen Verbesserungen für den Recall eines IRS angewandt (vs. dem damit verbundenen Abstieg der Precision) (vgl. Kowalski 1997, 67). Der Recall wird verbessert, da die unterschiedlichen morphologischen Varianten eines Suchbegriffs erfasst sind. Es ist wahrscheinlich, dass wenn ein User eine Anfrage mit “Spieler” startet, er auch an Dokumenten interessiert ist, die das Wort „Spiel“ beinhalten.

[70] engl. *performance*

Ein *Stemming*-Algorithmus kann auf dem linguistischen Wissen über die sprachspezifischen Suffixe (flexive und evtl. derivative), Endungen etc. basieren. Er kann aber auch mit rein statistischen Verfahren oder mit externen Datenbanken arbeiten. Nach der Methode der *Konflation* (Grundformenreduktion) können Algorithmen bzw. Stemmer in mehrere Gruppen eingeteilt werden:

1. *Affix removal Algorithmen* - entfernen die Präfixe, Suffixe und Endungen in einer bestimmten Reihenfolge und hinterlassen einen - manchmal auch etwas transformierten - Stem. Normalerweise wird die längste mögliche Sequenz entfernt, und das Verfahren wird wiederholt bis keine weiteren Regeln anzuwenden sind.
2. *Successor variety stemmer* - verwenden als Basis für das *Stemming* Frequenzen von Buchstabensequenzen im Text. Es wird von der Annahme ausgegangen, dass es sehr wahrscheinlich ist, dass am Ende eines zusammenhängenden Wortsegments die Zahl der möglichen Nachfolgebuchstaben abnimmt.
3. *Stemmer nach der n-gramm Methode* - konflatieren die Terme aufgrund ihrer gemeinsamen Bigramme oder n-gramme. Im folgenden Beispiel wurden die Wörter „Statistik“ und „statistisch“ in Bigramme aufgeteilt.

Statistik	=	st ta at ti is st ti ik
Bigramme:		st ta at ti is st ik (7)
statistisch	=	st ta at ti is st ti is sc ch
Bigramme:		st ta at ti is st sc ch (8)

Die Wörter „Statistik“ und „statistisch“ können in sieben, bzw. acht Bigramme unterteilt werden. Dabei sind Bigramme die Zweier-Einheiten, die sich nicht wiederholen. Die beiden Wörter haben 6 Bigramme gemeinsam. Mit diesen Angaben kann das *Ähnlichkeitsmaß S* berechnet werden :

$$Ähnlichkeitsmaß\ S = \frac{2 * (Anzahl\ der\ gemeinsamen\ Bigramme)}{(Bigramme\ 1.Wort) + (Bigramme\ 2.Wort)}$$

In diesem Beispiel wäre das *Ähnlichkeitsmaß* : $S = \frac{2 * 6}{7 + 8} = \frac{12}{15} = 0{,}8$.

4. ***Table lookup Stemmer*** - die Terme und ihre entsprechenden Stems sind alle in einer Tabelle gespeichert.

Term	Stem
Spiel spielerisch gespielt Spieler	spiel

Tab. 11: *Beispiel für table-lookup*

Sind die zu indexierenden Dokumente in verschiedenen Sprachen verfasst, so ist beim *Stemming* darauf zu achten, dass die Grammatik und die Sonderzeichen (z.B. die diakritischen Zeichen im Tschechischen) der jeweiligen Sprache mitberücksichtigt werden. Festzuhalten ist, dass die Sprachwerkzeuge (hier die *Stemmer*) sprachabhängig sind und lokalisiert werden müssen. Zum *Stemming* gibt es verschiedene Algorithmen für verschiedene Sprachen. Ein Problem würde auftauchen, wenn z.B. eine dem *Stemming* vorgelagerte Umwandlung des zu indizierenden Wortes beispielsweise in ASCII-Zeichen die diakritischen Zeichen eliminiert. Dies kann dem Wort eine völlig andere Bedeutung geben. So haben z.B. die tschechischen Wörter *ples* – „(Tanz-)Ball“ und *pleš* – „Glatze“, neben einer anderen Bedeutung, einen anderen Wortstamm und sind strikt voneinander zu unterscheiden.

Weitere Beispiele für die Bedeutung der diakritischen Zeichen wären:

může – „können“ (3. Person, Singular, Präsens) und
muže - „Mann“ (Singular, Genitiv)

Řek – „Grieche“ und
Rek – typischer Hundvorname (vgl. „Bello“ im Deutschen)

Ferner ist beim *Stemming* zu beachten, dass es auch vom Anwendungsbereich, genauer gesagt vom Korpus, abhängig ist. Ein weiteres Kriterium, das das *Stemming* beeinflusst, ist die Entscheidung, ob ein starkes oder ein schwaches *Stemming* bevorzugt wird. Diese Entscheidung hängt wiederum mit dem Recall/Precision-Problem zusammen. Denn, wie schon zu Anfang dieses Kapitels erwähnt, wird durch ein starkes *Stemming* der Recall eines IRS verbessert, da die unterschiedlichen morphologischen Varianten eines Suchbegriffs erfasst sind, was zugleich einen

Abstieg der Precision mit sich führt. So würde durch das *Stemming* die Anfrage mit bspw. dem Term "Spieler" auch Dokumente, die die Wörter „Spiel", „spielen" und „spielerisch" beinhalten, zurückliefern.

Durch das *starke* oder *schwache Stemming* kann es zu zwei unerwünschten Ergebnissen kommen: dem *Overstemming*, bzw. dem *Understemming.*[71]

Beim *Overstemming* wird eine zu lange Zeichenkette abgetrennt, was dazu führt, dass Wörter mit unterschiedlichen Bedeutungen und aus unterschiedlichen Wortfamilien auf ein und dieselbe Form reduziert werden und somit unerwünschterweise gleichgesetzt werden, z.B. wenn „Kommunismus" mit „Kommunikation" und „kommunizieren" gleichgesetzt wird, indem es auf den *Stem* „kommun" reduziert wird.

Dagegen wird beim *Understemming* eine zu kurze Zeichenkette abgetrennt. Dies führt dazu, dass unterschiedliche Wortformen mit ein und derselben Grund- oder Stammform wie unterschiedliche Wörter behandelt werden, z.B. sollten die Wörter "Kommunikation" und "kommunizieren" korrekterweise auf den gleichen *Stem* reduziert werden.

Die Entwicklung eines *Stemmers* ist eine experimentelle Wissenschaft, da Algorithmen nicht verifiziert werden können, sondern erst an Textkorpora und in der Praxis getestet werden müssen. Die so erzielten Ergebnisse können bspw. mittels der Speicher-, bzw. Laufzeit-Effizienz und der Retrieval-Effizienz (gemessen an den Recall- und Precision-Werten) evaluiert werden.

Im folgenden Kapitel wird anhand des Universal-Stemmers *EgoThor* die genaue Funktionsweise eines Stemming-Verfahrens vorgestellt, die wiederum von dem in dieser Arbeit analysierten polnischen Stemmer *STEMPEL* übernommen wurde.

[71] vgl. http://www.bui.haw-hamburg.de/pers/ursula.schulz/astep/le4_step_3.html

5.3.1 Der Universal-Stemmer *EgoThor*

Das EgoThor-Projekt ist ein in Java geschriebenes System für die Indexierung und anschließende Suche in Texten, das auch einen Stemmer beinhaltet. Die möglichen Anwendungsbereiche von EgoThor werden von den Entwicklern folgendermaßen beschrieben:

> "suitable for nearly any application that requires full-text search, especially cross-platform. It can be configured as a standalone engine, metasearcher, peer-to-peer HUB, and, moreover, it can be used as a library for an application that needs full-text search."

Das komplette System ist unter http://www.EgoThor.org im Internet frei verfügbar.

Der EgoThor-Stemmer ist ein Universal-Stemmer, d.h. seine Funktionsweise ist sprachunabhängig.[72] Er ist in der Lage, jede Sprache zu verarbeiten, indem er die Stemmingregeln der gegebenen Sprache anhand einer sprachspezifischen Beispieldatei erlernt.

Dies bedeutet wiederum, dass EgoThor durch das ihm zugrunde liegende Lernprinzip um jede Sprache beliebig erweiterbar ist. Bisher wurden elf europäischen Sprachen an ihm getestet (Dänisch, Deutsch, Englisch, Französisch, Italienisch, Niederländisch, Norwegisch, Portugiesisch, Russisch, Schwedisch und Spanisch). Die Erweiterung *EgoThors* um eine noch nicht vorhandene Sprache wird erreicht, indem eine Beispieldatei für diese Sprache angelegt wird und diese dann *EgoThor* für die Erstellung einer *Trie-Struktur* übergeben wird. Auf die Beispieldatei und die Beschreibung der Umwandlung in eine Trie-Struktur wird später in diesem Kapitel genauer eingegangen.

Weitere Schlüsseleigenschaften von EgoThor sind, dass er laut Aussagen der Entwickler so schnell wie ein einfacher Table-lookup-Stemmer ist und 50% weniger Speicherplatz benötigt als vergleichbare Systeme.

[72] Bei den Begriffen Universal-Stemmer und sprachunabhängig ist jedoch zu beachten, dass der Stemmer nur insofern universell, bzw. sprachunabhängig ist, dass er ausgehend von den sprachspezifischen Regeln für jede Sprache den gleichen Algorithmus verwendet.

Auf die Funktionsweise dieses Stemmers wird in den folgenden Abschnitten genauer eingegangen, da der in dieser Arbeit untersuchte polnische Stemmer *STEMPEL* auf *EgoThor* basiert. Die Beschreibung der Funktionsweise ist komplett dem Artikel „Semi-automatic stemmer evaluation“ von Leo Galamboš (Galamboš, *2004a)* entnommen.

Das Prinzip, das dem Stemmer zugrunde liegt, beruht auf der Annahme, dass die Prozesse des *Stemming* von der Struktur her für jede Sprache gleich sind. Beim *Stemming* werden bestimmte Umformungsregeln ausgeführt. Diese sog. *Stemmingregeln* sind sprachspezifisch und werden durch die einmalige Verarbeitung einer Beispieldatei (vom Entwickler, Leo Galamboš, „dictionary“ genannt) erfasst. Die Beispieldatei enthält die Wortumformungen, die sich aus der Grundform eines Wortes (im folgenden Beispiel `abermal)` und seinen Variationen zusammensetzt:

```
abermal  abermalige abermaligen abermals.
```

Anhand der Beispieldatei lernt *EgoThor* die Stemmingregeln der jeweiligen Sprache. Dazu werden für jeden Eintrag die Grundform und die dazugehörige Umformungsabfolge, die das Wort in seine Grundform überführt, abgespeichert.

Die Umformungsabfolge wird als patch command (P-command) bezeichnet und kann folgende auszuführende Operationen enthalten:

- „removal“- einen Buchstaben entfernen,
- „insertion“ - einen Buchstaben hinzugen,
- „substitution“ - einen Buchstaben ersetzen oder
- „no operation“ – das Wort unverändert lassen.

Diese Umformungsoperationen werden auf einem Wort von rechts nach links angewendet und zusammen mit der Grundform in einem *Trie* abgespeichert. Ein *Trie* (abgeleitet aus dem engl. re*Trie*val)[73] ist eine baumartige Datenstruktur, um Wortformen oder Teile von solchen, z.B. die Endungen, so abzuspeichern, dass der Zugriff darauf sehr schnell geschehen kann. Ein Pfad der Baumstruktur geht von der

[73]Der Begriff *Trie* wurde von E. Fredkin 1960 in seinem Artikel *Trie Memory* eingeführt (erschienen in: Communications of the ACM, 3(9):490-499).

Wurzel bis zum Blatt und stellt eine abgespeicherte Zeichenfolge (*String*) dar, dabei wird jedem Knoten ein Buchstabe zugeordnet. Der Endbuchstabe befindet sich in der Wurzel.[74] Wenn ein Wort gestemmt werden soll, dann wird es über dessen Endung im Trie gesucht. Im Falle des EgoThor-Stemmers werden in den Blättern die Umformungsregeln mit abgespeichert.

Wenn alle Einträge der Beispieldatei in den Trie eingefügt worden sind, ergibt sich eine sehr große Datenstruktur, die redundante Informationen enthält. In Abb. 11 ist das die zweimal vorkommende Stemmingregel der Strings „abilities“ und „activities“. Aus diesem Grund werden überflüssige Informationen gelöscht und die Regeln verallgemeinert. Wird dies auf dem eben genannten Beispiel angewandt, entsteht folgender reduzierter Trie:

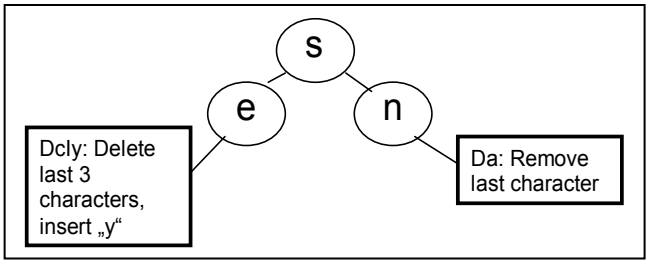

Abb.14: *Reduzierte Trie-Struktur nach Galamboš 2004*

Die verbliebenen zwei Regeln lauten: „Endet ein Wort auf „es“, dann entferne die letzten drei Buchstaben und füge ein „y“ ein.“ Und „Endet ein Wort auf „ns“, dann entferne den letzten Buchstaben.“

Diese optimierte Datenstruktur benötigt wenig Speicherplatz und kann schnell durchsucht werden. Die Verwendung der Beispieldatei kombiniert zwei verschiedene Stemming-Strategien. Zum einen werden die Vorzüge der table look-up Methode verwendet, die das korrekte Verarbeiten von irregulären Umformungen ermöglicht und zum anderen wird durch die Affix-Abtrennung keine komplexe Datenstruktur benötigt (vgl. Galamboš, 2004).

Der Stemm-Algorithmus ist für alle Sprachen gleich, denn für jede Sprache wird eine eigene Trie-Struktur mit den spracheigenen Umformungsregeln abgelegt, die sich nur

[74] vgl. http://www.uni-trier.de/uni/fb2/ldv/ldv_wiki/index.php/Trie

in ihrem Inhalt unterscheidet. Somit bleibt EgoThor sprachunabhängig und die universale Funktionalität des Stemmers wird erreicht.

5.3.2 Der polnische Stemmer *STEMPEL*

STEMPEL ist ein Stemmer für die polnische Sprache, erstellt von Andrzej Bialecki.[75] Der Quellcode für den STEMPEL wurde nahezu unverändert von dem in Kapitel 5.3.1 vorgestellten EgoThor Projekt übernommen. Hinzugefügt wurde die Stemming Table für Polnisch, aus der der Stemmer seine Umformungsregeln generiert. Das komplette Java-Package ist unter der Apache-Style Open Source Lizenz verfügbar und kann unter http://getopt.org/stempel heruntergeladen werden.

Der polnische Stemmer STEMPEL benutzt genauso wie EgoThor eine algorithmische Methode. Der Hauptvorteil von algorithmischen Stemmern liegt darin, dass sie in der Lage sind vorher noch nicht gesehene Wortformen mit einer hohen Genauigkeit zu verarbeiten.

Auf der beigefügten CD befindet sich STEMPEL in Form eines Eclipse-Projektes. Im Falle, dass der Quellcode des Stemmers STEMPEL heruntergeladen wird, wird *Jakarta Ant* benötigt, um das Jar-Verzeichnis zu erstellen. Dieses auf JAVA basierende Tool kann unter http://ant.apache.org/ heruntergeladen werden.

Weiterhin sind folgende Dependenzen zum Laufen von STEMPEL notwendig:

- *JRE System Library [jre.1.5.0]*
- *lucene-1.4.3.jar*
- *junit.jar*,[76]
- Mit der Klasse *AnalyzerDemo.java* kann STEMPEL getestet werden.

Der polnische Stemmer STEMPEL wurde in dieser Arbeit herangezogen, da zu diesem Zeitpunkt kein Stemmer für die tschechische Sprache vorhanden war. Es hätte die Möglichkeit bestanden, EgoThor durch das ihm zugrunde liegende Lernprinzip um Tschechisch zu erweitern. Dies hätte jedoch bedeutet, dass er die sprachspezifischen

[75] weitere Informationen zu dem Autor unter : http://getopt.org
[76] Version 3.8.1

Stemmingregeln durch eine Beispieldatei erfassen müsste. Das Erstellen einer solchen Beispieldatei hätte sich für diese Arbeit zu umfangreich gestaltet.

Ein weiterer Grund für die Verwendung eines polnischen Stemmers liegt darin, dass durch den nahen Verwandtschaftsgrad der polnischen und tschechischen Sprache (vgl. Kapitel 3) eine ähnliche Flexions- und Derivationsmorphologie vermutet wurde. Laut (Bußmann 1990, 706) wird

> „die Zugehörigkeit von Sprachen zu einer Sprachfamilie in der Regel durch phonologische, morphologische und lexikalische Übereinstimmungen erwiesen".

In dieser Arbeit wurde getestet, ob die vorhandenen Übereinstimmungen der polnischen und tschechischen Sprache für ein vergleichbares Stemming-Verhalten ausreichend waren.

Das anschließende Kapitel präsentiert die Ergebnisse, die durch die Anwendung von STEMPEL auf tschechische Texte erzielt wurden und beinhaltet eine Diskussion über eine Verwendung des Stemmers STEMPEL im Hinblick auf MIMOR@CLEF.

5.3.3 Auswertung der Ergebnisse im Hinblick auf MIMOR@CLEF

Bevor die Ergebnisse des polnischen Stemmers STEMPEL ausgewertet werden können, muss zunächst das Auswahlverfahren für die zu stemmenden Terme beschrieben werden. Bei der Auswahl der zu stemmenden Terme wurden zwei Kriterien beachtet:

1. Die zu stemmenden Terme sind entweder Wörter, die zu der gleichen Wortfamilie gehören, wie z.B.: *adaptace, adapter* und *adaptovat* (dt. „Adaptation", „Adapter" und „adaptieren")
2. oder umfassen verschiedene Konjugationsformen eines Verbs, z.B.: *číst, čtu, četl, četla, čti, čet, čtěte* (dt. „lesen", „lese", „las", „las", „lies", „gelesen", „lest").

Diese zwei Kriterien führten zu günstigen Vorraussetzungen für die Auswertung der Ergebnisse: zum einen wurden die zu stemmenden Terme gruppiert (bspw. wurde eine Wortfamilie zu einer Gruppe zusammengefasst), zum anderen konnte auf diese Weise die Reduktion auf einen *Stem* pro Gruppe erwartet und leichter überprüft werden.

Insgesamt wurden in zwei Durchläufen 73 Terme von STEMPEL gestemmt. Der erste Durchlauf, der das Stemmen von 40 Termen umfasste, zeigte bereits, dass die Stemming-Prozedur nicht, wie erhofft, positiv ausfällt. Die zu stemmenden Terme und die von STEMPEL generierten Stems sind für die ersten 18 Terme in der folgenden Tab. 12 aufgeführt. Eine vollständige Auflistung befindet sich im Anhang.

	Term	erwarteter Stem	Stem
1	adaptace	adapt	adaptaka
2	adaptér	adapt	adaptéra
3	adaptovat	adapt	adaptovata
4	balík	balí	balík
5	balít	balí	balít
6	celkem	celk	celkem
7	celkový	celk	celkový
8	celkově	celk	celkově
9	cvičení	cvič	cvičení
10	cvičený	cvič	cvičený
11	cvičit	cvič	cvičit
12	cvička	cvič	cvička
13	cvičný	cvič	cvičný
14	dědic	dědi	ić
15	dědictví	dědi	dědictví
16	dědičnost	dědi	dědičnosta
17	dědičný	dědi	dědičný
18	dědit	dědi	dědit

Tab. 12: *Ergebnistabelle von STEMPEL für 18 tschechische Terme des ersten Durchlaufs*

Bei 28 von 40 Termen war der generierte Stem identisch mit dem zu stemmenden Term. Der Stemmer hatte die Termform unverändert gelassen. Lediglich in einem Fall wurde der zu stemmende Terme auf eine potentiell korrekte Form verkürzt:

`individualita → individualit`

Bei sechs Termen führte das Stemmen zu einer Verlängerung des Terms, wie bspw. in den Linien 2 und 16 der Tab. 12 zu sehen ist.

Der zweite Durchlauf mit 33 Termen beschränkte sich auf das Stemmen von regelmäßigen Verben. Er beruhte auf der Annahme, dass diese durch ihre Ähnlichkeit eher korrekt gestemmt werden würden, als die Terme im ersten Durchlauf. Die ausgewählten tschechischen Verben waren intellektuell eindeutig einem Stem zuzuordnen. Gleichzeitig sollte der zweite Durchlauf die bisherigen Ergebnisse entweder bestätigen und die bisherigen Ergebnisse bekräftigen oder zeigen, dass zumindest die regelmäßigen Verben besser gestemmt werden. Die Ergebnisse dieses Durchlaufs befinden sich ebenfalls im Anhang.

Folgendes Beipiel zeigt anhand des Verbes *koupit* (dt. „kaufen) exemplarisch die Stemmingergebnisse für regelmässigen Verben:

	Term	erwarteter Stem	Stem	korrekt	unverändert
1	koupit	koup	koæ		
2	koupím	koup	koupím		x
3	koupíš	koup	koupí[1]		
4	koupí	koup	koupí		x
5	koupíme	koup	koupímy		
6	koupíte	koup	koupíte		x
7	koupějí	koup	koupìjí		
8	koupila	koup	koupil		
9	koupil	koup	koupil		x
10	kup	kup	kup	(x)	x
11	kupte	kup	kupte		x

Tab. 13: *Stemming-Ergebnisse von STEMPEL für regelmäßige Verben am Beispeil des Verbes „koupit“*

Das in Tab. 13 angeführte Beispiel enthält in Zeile 11 den einzigen Fall, dass „korrekt“ gestemmt wurde. Jedoch kam wahrscheinlich dieses Ergebnis zustande, weil zwei Faktoren zusammentrafen: zum einen stimmt der erwartete Stem mit dem ursprünglichen Term überein und zum anderen wurde durch STEMPEL keine Veränderung vorgenommen. Aus diesem Grund ist das korrekte Ergebnis eingeklammert und das hier als gültig betrachtete Ergebnis lautet in diesem Fall *unverändert*.

Abb. 15 hält die Ergebnisse für den zweiten Durchlauf fest:.

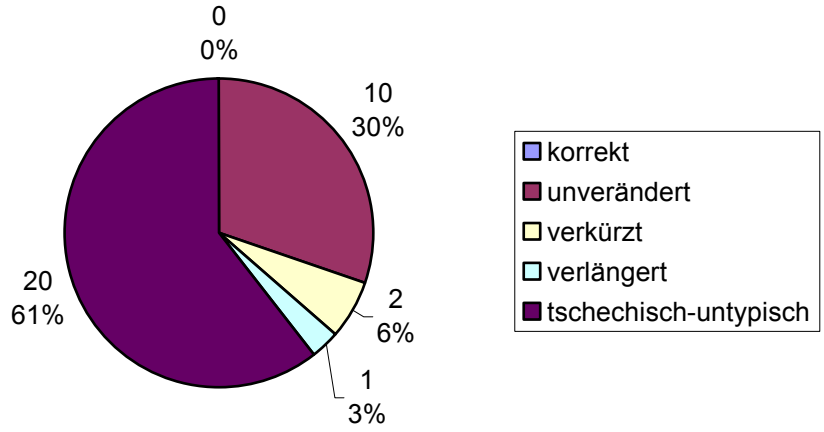

Abb. 15: *Ergebnismengen für den zweiten Durchlauf von STEMPEL*

Von den insgesamt 33 Termen wurde kein einziger auf den erwarteten Stem reduziert. Der Wert für die korrekt gestemmten Terme ist Null. In zwei Fällen löschte der Stemmer den Endbuchstaben. In der Abb. 12 sind diese Ergebnisse unter *verkürzt* zusammengefasst. Ein Stem wurde statt reduziert, um einen Endbuchstaben erweitert. STEMPEL lieferte für 10 Terme unveränderte Formen zurück. 20 der vom ihm erstellten Stems wiesen für Tschechisch untypische Eigenschaften auf. In der Tab. 12 ist das bspw. die Form „`ić`“. Diese untypischen Wortausprägungen kamen vermutlich durch folgende Gründe zustande:

- Die Flexions- und Derivationsmorphologie der polnischen und der tschechischen Sprache sind zu unterschiedlich, sodass der für die polnische Sprache erstellte Stemmer STEMPEL bei den Termoperationen auf den tschechischen Input auf tschechische Wortformen traf und diese in für die polnische Sprache typische Ausprägungen umgewandelt wurden.

- Teilweise generierte STEMEPL durch das Verarbeiten von tschechischem Input auch für die polnische Sprache untypische Wortausprägungen, wie bspw. in der Tab. 13 Linie 1 zu sehen ist: `koæ`. Das Zeichen „æ“ kommt im Polnischen nicht vor.

Diese Umwandlung beruht vermutlich auf den unterschiedlichen Zeichensatz, die den beiden Sprachen zugrunde liegen. Keines der tschechischen Sonderzeichen ist im Polnischen vorhanden (und umgekehrt). Weiterhin existieren in der tschechischen Sprache weitaus mehr Sonderzeichen als im Polnischen (30 vs. 18).

Für die Verwendung von STEMPEL im Hinblick auf eine Anwendung auf tschechischen Input ergeben sich folgende Konsequenzen:

- Ob die von STEMPEL für die tschechische Sprache erstellten Stems sich im Hinblick auf den IR-Prozess als positiv herausstellen, ist an dieser Stelle nicht zu beurteilen. Die endgültige Entscheidung, ob diese Stems sich als sinnvoll erweisen, kann nur in zukünftigen Tests im IR-Prozess ermittelt werden.

- Eventuell stellen die so generierten Stems eine Verbesserung zum Nicht-Stemmen dar. Auch dies müsste in einem vergleichenden Test ermittelt werden.

- Eine Variante und mögliche Lösung für das Stemming-Problem der tschechischen Sprache könnten bspw. statistische Stemming-Verfahren darstellen.

5.4 Intellektuell erstellter Text-Katalog für Tschechisch

Ein weiteres Ergebnis dieser Arbeit stellt der intellektuell erstellte Text-Katalog für die tschechische Toplevel-Domain von WebCLEF dar. Dieser Text-Katalog bestimmt eindeutig die Sprachen von 700 HTML-Dokumenten aus den tschechischen Domains cz/001.gz, cz/002.gz, cz/003.gz, cz/004.gz, cz/005.gz und cz/006.gz. Er befindet sich auf der beigefügten CD.

In diesem Kapitel wird zunächst die WebCLEF-Dokument-Kollektion beschrieben. Anschließend erfolgt die Schilderung der Vorgehensweise für die Erstellung des

intellektuellen Text-Katalogs. Die Ausgangsbasis stellten dabei die vom automatischen Sprachidentifizierer generierten Ergebnisse, die in einem weiteren Schritt analysiert wurden. Im Vordergrund stand dabei, die Faktoren zu determinieren und zu klassifizieren, die zu Fehlern bei der automatischen Sprachidentifikation geführt haben. Diese Faktoren wurden im intellektuell erstellten Text-Katalog mittels Kommentaren festgehalten.

CLEF 2005 bietet eine Serie von acht Aufgabenstellungen, sog. *Tracks*, um die verschiedenen Aspekte in IR-Systemen zu evaluieren. Das *Multilinguale Web Track* (auch WebCLEF genannt) wird von der Universität Amsterdam koordiniert.[77]

Bei der WebCLEF-Initiative werden cross-linguale IR-Systeme in einer Web-Umgebung evaluiert. Diese entspricht der natürlichen Umgebung für das multi-, bzw. cross-linguale Retrieval. Gerade in Europa erfolgt die Suche nach Informationen im Web vor allem multilingual. Diese Informationen stammen aus den verschiedensten Bereichen: von Wirtschaft, Recht, über Kultur, Lehre bis hin zu Freizeit- und Reiseangeboten. Die Edition 2005 von WebCLEF hat zwei *Main Tasks: gemischt monolingual* und *multilingual.*

Die WebCLEF Dokument-Kollektion setzt sich zusammen aus ca. 3,9 Mio. Webseiten von europäischen Regierungen des EuroGOV-Korpus. Der Korpus besteht laut des Berichtes der Universität Amsterdam vom 1. März 2005[78] aus 157 Dateien, die jeweils maximal 25000 HTML Dokumente beinhalten. Gezipt beträgt die Größe des Korpus 11GB, entzipt ca. 84GB.

Die Tschechische Republik gehört mit 320000 Seiten zu den fünf „Ländern“[79] mit der höchsten Anzahl an Seiten. Die anderen vier sind: Finnland mit 660000, Deutschland mit 450000, EU mit 375000 und Ungarn mit 330000.[80]

[77] http://ilps.science.uva.nl/WebCLEF/
[78] http://lit.science.uva.nl/Teaching/0405/IIResources/ii-0405-week04-2-8up.pdf
[79] Der Begriff “Länder” ist nicht stellvertretend für alle Toplevel-Domains, da auch die Toplevel-Domain der EU, `eu.int`, vertreten ist. Diese umfasst alle 20 offiziellen Sprachen der EU.
[80] WebCLEF (2005) Participation Guidelines

Die EuroGOV-Kollektion umfasst 27 Toplevel-Domains, die unterteilt werden in 13 *Main Domains* und 14 *Additional Domains*. Die folgende Tab. 14 führt diese auf.

EUROGOV Collection Domains.			
Main domains		Additional Domains	
Domain	Country	Domain	Country
.cz	Czech Republic	.at	Austria
.de	Germany	.be	Belgium
.es	Spain	.cy	Cyprus
.eu.int	European Union	.dk	Denmark
.fi	Finland	.ee	Estonia
.fr	France	.gr	Greece
.hu	Hungary	.ie	Ireland
.it	Italy	.lt	Lithuania
.nl	The Netherlands	.lu	Luxemburg
.pt	Portugal	.lv	Latvia
.ru	Russia	.mt	Malta
.se	Sweden	.pl	Poland
.uk	United Kingdom	.si	Slovenia
		.sk	Slovakia

Tab. 14: *Toplevel-Domains von EuroGOV*

Ferner ist die EuroGOV-Kollektion in Verzeichnisse unterteilt, sodass jeweils ein Verzeichnis einer Toplevel-Domain zugeordnet ist. Jedes Verzeichnis enthält ein oder mehrere gepackte Dateien. Jede Datei enthält wiederum bis zu 25000 Dokumente.

Den Verzeichnissen liegt folgendes Format zugrunde:

```
<EuroGOV:bin
 domain=""       <!-- The top level domain -->
 id="">          <!-- The name of the file -->
<EuroGOV:doc
 url=""          <!-- URL of the page -->
 id=""           <!-- DocID of the format Exx-yyy-z -->
                 <!--  E is E and stands for EuroGOV -->
                 <!--  xx is the top level domain -->
                 <!--  yyy is the file name -->
                 <!--  z is the character offset of the document -->
 md5=""          <!-- MD5 checksum of the content of the page -->
 fetchDate=""    <!-- Fetch date of the page -->
 contentType=""> <!-- contentType as given by the web server -->
<EuroGOV:content>
<![CDATA[
... content ...  <!-- This is the actual page -->
]]>
</EuroGOV:content>
</EuroGOV:doc>
...
</EuroGOV:bin>
```

Es folgt ein Auszug aus der tschechischen Domain cz/005.gz:

```
1     <EuroGOV:bin domain="cz" id="005">
2     <EuroGOV:doc
3     url="http://portal.gov.cz/wps/portal/.cmd/cps/.c/320/.ce/6131
4     /.ps/X/_s.155/692/_mx.2701/6305/_s.155/692/_ps.6305/X/_ps.5104/M"
5     id="Ecz-005-35"
6     md5="4626a10c6089d50ebcea95ef48eb6846"
7     fetchDate="Tue Oct 19 16:11:13 MEST 2004"
8     contentType="text/html; charset=UTF-8">
9     <EuroGOV:content>
10    <![CDATA[
11    <!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
12    <html>
13      <head>
14        <title>Portal verejne spravy</title>
15      <meta http-equiv="content-language" content="cs">
16      <meta http-equiv="content-type" content="text/html;   charset=UTF-8">
17      <meta name="author" content="all: Ministerstvo Informatiky;
18       e-mail: posta@micr.cz">
```

Jedes Dokument, bzw. jede Webseite, ist mit einer ID versehen. In einigen Fällen kommt es vor, dass verschiedene IDs auf die gleiche Seite verweisen. Im vorangegangenen Beispiel war die ID der Seite: id="Ecz-005-35" (siehe Zeile 5). Bei der Auswahl der Seiten deren Sprache in dieser Arbeit bestimmt werden sollte, wurden Doppelungen, so weit es ging, vermieden. Ganz auszuschließen sind diese aber aufgrund der großen Menge nicht. Weiterhin wurden folgende HTML-Dokumente bei der intellektuellen Identifizierung außer acht gelassen:

- HTML-Dokumente, die nur aus Bildern bestanden,
- Fehlermeldungen lieferten
- oder nur einen Link zum Runterladen bspw. eines DOC oder PDF-Dokumentes beinhalteten.

Das Gegenstück zur intellektuellen Sprachidentifizierung stellt die automatische Sprachidentifizierung dar. Sie beruht auf der Tatsache, dass im Web die Sprachen der HTML-Dokumente a priori nicht bekannt sind. In dieser Arbeit wurde der automatische Sprachidentifizierer der WebCLEF Organisatoren evaluiert. Die

Evaluierung dieser Ressource liefert wichtige Erkenntnisse über die Faktoren, die die Bestimmung der Sprache eines Dokumentes beeinflussen.

Die Ergebnisse des automatischen Sprachidentifizierers der WebCLEF Organisatoren lagen in einer Datei, dem sog. Text-Katalog `cz_TextCat.txt` vor. In diesem Text-Katalog waren den einzelnen HTML-Seiten die Sprache zugeordnet, in der das Dokument verfasst war. Das folgende Beispiel zeigt eine eindeutige Zuordnung der Sprache Tschechisch für die Seite mit der ID Ecz-001-48227:

```
Ecz-001-48227 :: czech-iso8859_2
```

In den meisten Fällen wurden den Seiten mehrere Sprachen zugeordnet, da der automatische Sprachidentifizierer nicht eindeutig die Sprache bestimmen konnte, sodass in den meisten Fällen folgende mehrdeutige Ergebnisse zustande kamen:

```
Ecz-001-1009648 :: slovak-windows1250 or czech-iso8859_2 or slovak-ascii
```

Die intellektuelle Sprachidentifikation ermöglicht ein eindeutiges Ergebnis, das in dem eben angeführten Beispiel Tschechisch als Sprache des Dokumentes bestimmt. Dazu wurde die ID der Seite aus dem automatisch erstellten Text-Katalog in dem Verzeichnis der Toplevel-Domain gesucht und mittels der dazugehörigen URL die Seite aufgerufen. Nach genauerer Untersuchung der Seite wurde in dem intellektuell erstellten Text-Katalog für dieses Dokument folgender Eintrag erstellt:

```
Ecz-001-1009648 :: czech.
```

Welche Faktoren die automatische Bestimmung der Sprache beeinflussen, wird etwas später in diesem Kapitel erläutert. Zunächst wird auf die Zusammenstellung der untersuchten Toplevel-Domains eingegangen.

Einführend muss gesagt werden, dass fast alle Dokumente der tschechischen Toplevel-Domains auf Tschechisch verfasst sind, da die angegebenen URLs in den HTML-Dokumenten auf Seiten des Portals der tschechischen Regierung verwiesen. Folgende Graphik Abb. 16 veranschaulicht die Sprachverteilung innerhalb der tschechischen Toplevel-Domains:

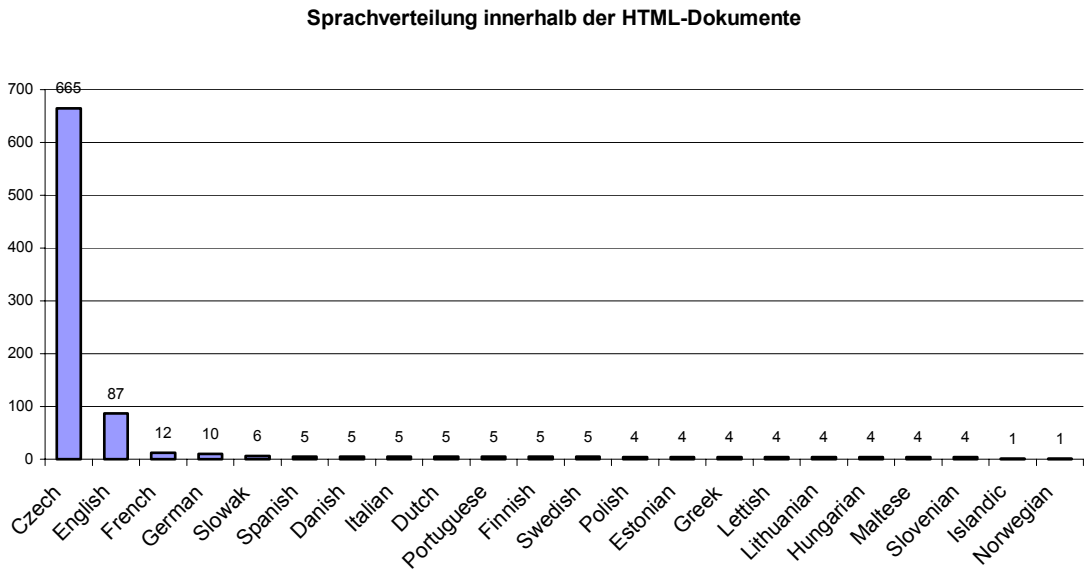

Abb.16 : *Sprachverteilung innerhalb der HTML-Dokumente der untersuchten tschechischen Toplevel-Domains*

Insgesamt wurden in den 700 HTML-Dokumenten 22 Sprachen intellektuell identifiziert. An dieser Stelle ist anzumerken, dass auch fremdsprachige Passagen innerhalb eines tschechischen Textes intellektuell erfasst wurden. Aus diesem Grund ist die Graphik Abb. 16 folgendermaßen zu interpretieren: 665 Dokumente von 700 waren auf Tschechisch verfasst oder enthielten tschechischsprachige Passagen. Diese Passagen konnten Zitate, Überschriften oder bspw. tschechische Adressen mit Firmennamen sein. Fremdsprachige Passagen wurden in Form eines Kommentars festgehalten. Für einen in Tschechisch verfassten Text mit einer englischsprachigen Passage ergab sich folgender Eintrag im intellektuell erstellten Text-Katalog:

```
Ecz-001-1800762 :: czech (in last chapter english firm names)
```

War das Dokument in einer anderen Sprache als Tschechisch verfasst, so wurde auch diese Sprache bestimmt. Das folgende Dokument ist auf Englisch verfasst und erhält somit folgenden Eintrag:

```
Ecz-001-49891703 :: english
```

In wie weit mehrere Sprachen innerhalb eines Dokumentes vorkommen, also wie oft fremdsprachige Passagen innerhalb eines Textes auftreten, illustriert das Diagramm in der Abb. 17.

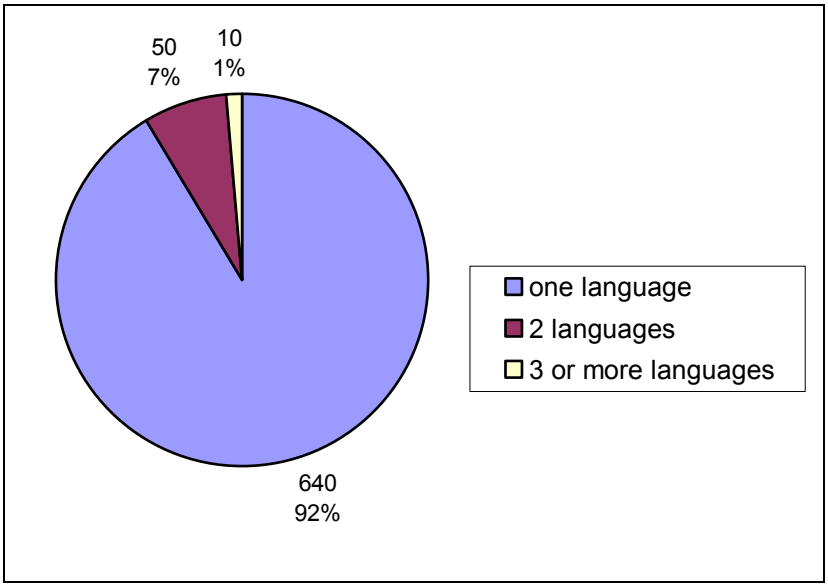

Abb. 17 : *Sprachen innerhalb eines Dokumentes*

Der größte Teil der Dokumente (92%) war ausschließlich in einer Sprache verfasst. 50 der insgesamt 700 untersuchten Dokumente enthielten neben der Hauptsprache des Dokuments[81], eine weitere Sprache in Form von Passagen und nur in 10 Dokumenten war es der Fall, dass drei oder mehrere fremdsprachige Passagen innerhalb des Textes erschienen. Beispiele hierfür sind:

```
Ecz-002-64908330 :: czech, french, english, danish, german, spanish, italian,
                    dutch, portuguese, finnish, islandic, norwegian, Swedish

Ecz-002-68507758 :: english, czech, french, german
```

Nun soll erläutert werden welche Faktoren die automatische Bestimmung der Sprache beeinflussen. Der entscheidende Aspekt an dieser Stelle ist, neben der Anzahl der Sprachen innerhalb des Dokuments, die Länge der Dokumente.[82] Allgemein kann

[81] damit ist die Sprache gemeint, in der das Dokument zum größten Teil verfasst ist.
[82] Weitere Faktoren, die die Bestimmung der Sprache beeinflussen sind das Anzeigen von diakritischen Zeichen und Fehlermeldungen, die sich innerhalb der Seite aufbauen. Auf diese Faktoren wird in dieser Arbeit aber nicht weiter eingegangen, da diese nur vier Mal, bzw. ein Mal in der hierfür durchgeführten Studie auftraten.

gesagt werden, dass je länger ein Dokument ist, desto höher ist die Wahrscheinlichkeit, dass die durch den automatischen Sprachidentifizierer bestimmte Sprache korrekt ist.[83]

Die Dokumente wurden je nach Länge in folgende Kategorien eingeteilt: *normal*, *short* und *very short*. Die Länge *short* bezeichnet ein Dokument, das nur aus ca. 50 bis 100 Wörtern besteht. Dokumente mit einer Wortanzahl unter 50 wurden als *very short* eingestuft. Das Schaublid Abb. 18 zeigt die Verteilung der Dokumente innerhalb der untersuchten tschechischen Toplevel-Domains je nach Länge:

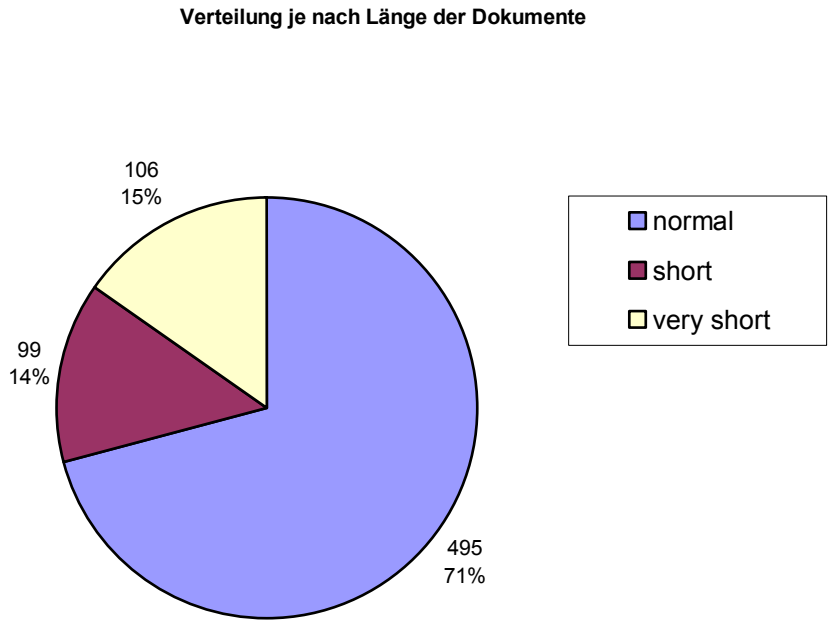

Abb. 18 : *Verteilung der Dokumente je nach Länge*

Aus dem Diagramm ist zu entnehmen, dass der Hauptanteil der Dokumente eine *normale* Länge aufweist und nur 14% bzw. 15% als *short* oder *very short* eingestuft wurden. Der Zusammenhang zwischen der Länge und den Fehlern bei der automatischen Sprachidentifizierung wird erläutert, nachdem geklärt wird, was Fehler in diesem Kontext darstellen.

[83] Was mit korrekt gemeint ist, wird im anschließenden Abschnitt deutlich, wenn die verschiedenen Stufen von Fehler dargestellt werden.

Bei der Auswertung der Ergebnisse wurde zwischen verschiedenen Stufen von Fehlern unterschieden:

- unpräzise
- fremdsprachige Passage nicht erkannt
- unbekannt
- falsch

Unter *unpräzise* ist zu verstehen, dass mehrere mögliche Sprachen für ein Dokument angegeben wurden, wobei die korrekte Sprache in der Auflistung enthalten war. Zum Beispiel die Angabe

```
Ecz-001-1226306 :: slovak-ascii or slovak-windows1250 or czech-iso8859_2 or english or danish or polish
```

für ein Dokument, das auf Tschechisch verfasst ist .

Die Anmerkung *fremdsprachige Passage nicht erkannt* bedeutet, dass nur eine Sprache angegeben wurde (in der Regel die Hauptsprache) und eine oder mehrere fremdsprachige Passagen nicht identifiziert wurden. Im folgenden Beispiel wurde nur Tschechisch erkannt:

```
Ecz-002-65574417 :: czech-iso8859_2
```

und nicht die englischsprachigen Passagen. Der korrekte Eintrag lautet:

```
Ecz-002-65574417 :: 1. half of the text in czech, 2. half in english, titles in both languages.
```

Die Angabe *unbekannt* bedeutet, dass der automatische Sprachidentifizierer nicht die Sprache bestimmen konnte und als Ergebnis *unknown* lieferte.

Eine Bestimmung der Sprache wurde als *falsch* eingestuft, wenn entweder nur ein falsches Ergebnis angegeben wurde (z.B. Slowakisch anstatt Tschechisch) oder wenn mehrere mögliche Sprachen angegeben wurden und die tatsächliche Sprache nicht in der Aufzählung vorhanden war (z.B. im Falle von `Ecz-002-68943601 :: portuguese or spanish` anstatt `Ecz-002-68943601 :: czech (very short)`).

In der Graphik Abb. 19 wurden die Ergebnisse zusammengetragen.

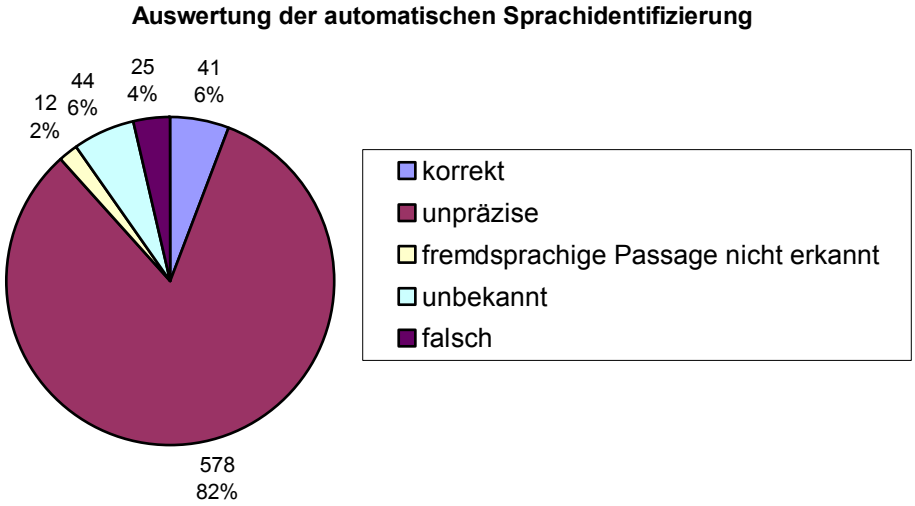

Abb. 19: *Auswertung der Ergebnisse des automatischen Sprachidentifizierers*

Auffällig ist, dass der Anteil der Dokumente deren Sprache *korrekt* eingestuft wurde, mit 6% sehr klein ausfällt. Jedoch sollten die als *unpräzise* eingestuften Ergebnisse nicht komplett als negativ bewertet werden, da diese zumindest die zutreffende Sprache in der Auswahl enthielten.

Weitere Untersuchungen haben gezeigt, dass die Länge eines Dokuments bei der automatischen Sprachindentifizierung von großer Bedeutung ist. Von den 106 Dokumenten, die als *very short* eingestuft wurden, lieferten 95 ein unpräzises Ergebnis, 4 ein *falsches*, 5 die Meldung *unknown* und bei 3 wurden die *fremdsprachigen Passagen* nicht erkannt. Lediglich in 5 Fällen lieferte der automatische Sprachidentifizierer ein korrektes Ergebnis für diese Dokumentenlänge. Die Untersuchungen bei Dokumenten, die als *short* eingestuft wurden, spiegelten eine ähnliche Verteilung wieder.

Zusammengefasst können die Faktoren, die die automatische Sprachidentifizierung negativ beeinflussen folgendermaßen klassifiziert werden:

- *Sprachverteilung* - Die Sprachverteilung beschreibt sie, ob das jeweilige Dokument in einer oder mehreren Sprachen verfasst ist. Fremdsprachige Passagen wurden entweder nicht erkannt oder führten zu einem unpräzisen Ergebnis.
- *Länge der Dokumente* – Je länger ein Dokument ist, desto höher ist die Wahrscheinlichkeit, dass die Sprache des Dokuments korrekt bestimmt wird.

Die Evaluierung des automatischen Sprachidentifizierers und der darauf aufbauende intellektuell erstellte Text-Katalog für die tschechische Toplevel-Domain von WebCLEF bieten einen Standard für die automatische Sprachidentifikation. Dadurch dass im WebCLEF-Kontext die jeweiligen Sprachen, in denen ein HTML-Dokument verfasst ist, nicht bekannt sind, wird eine richtige und eindeutige Sprachzuordnung erforderlich.

Mit der Problematik des Web Retrievals befasst sich auch die Magisterarbeit von Niels Jensen (2005).[84] Für seine Untersuchungen verwendet er das bereits in dieser Arbeit vorgestellte EuroGOV-Korpus.

Auch die gemeinsame Magisterarbeit von Olga Artemenko und Margaryta Shramko (2005)[85], die zum Ziel hat ein Werkzeug zur Sprachidentifikation zu entwickeln, geht in die gleiche Richtung. Ihr automatischer Sprachidentifizierer soll in der Lage sein, fremdsprachige Passagen innerhalb eines Textes zu erkennen.

Gemeinsam mit den eben genannten Arbeiten entsteht durch den hier erstellten Text-Katalog eine Basis, für weitere Projekte im Bereich des Web-Retrievals.

[84] „Multilinguales Web Retrieval am Beispiel des EuroGOV Korpus“
[85] „Entwicklung eines Werkzeugs zur Sprachidentifikation in mono- und multilingualen Texten“

Kapitel 6
Abschlussbetrachtung und Ausblick

Diese Arbeit gibt einen Einblick über die aktuelle Position und Möglichkeiten der tschechischen Sprache im IR. Es wurden informationslinguistische Ressourcen für Tschechisch vorgestellt, analysiert und erstellt. Ausgangspunkt für die vorliegende Arbeit war zum einen die Teilnahme an der Evaluierungsinitiative CLEF im Bereich WebCLEF und zum anderen die zukünftige Erweiterung von MIMOR um die tschechische Sprache.

Während der Untersuchung des aktuellen Stands der informationslinguistischen Ressourcen für die tschechische Sprache, zeigte sich, dass in verschiedenen Bereichen noch Forschungs- und Entwicklungsbedarf besteht. So war z.B. zum Zeitpunkt dieser Arbeit kein Stemmer für Tschechisch vorhanden.

Mit Hilfe des Korpus SYN2000 wurde eine Stoppwortliste für tschechische Textkorpora erstellt. Bei der Erstellung wurden einige hypothetische Entscheidungen getroffen, die auf vorigen Studien mit anderen Sprachen beruhten, z.B. die Aufnahme von Termen in die Stoppwortliste, die nicht zu den 200 häufigsten Wörtern des Korpus zählten. Diese Stoppwortliste erzielt im Korpus SYN2000 eine Abdeckung von fast 30% und kann in zukünftigen Tests an die jeweilige Domäne des Korpus angepasst werden.

Die Analyse einer möglichen Anwendung des polnischen Stemmers STEMPEL hat die in Kapitel 3.1 determinierten Besonderheiten und Schwierigkeiten belegt. Für Tschechisch bedarf es Ressourcen, die an die Eigenschaften der Sprache, d.h. die komplexe Flexions- und Derivationsmorphologie und den Zeichensatz, angepasst sind. In zukünftigen Tests sind die mögliche Verwendung und der definitive Einsatz von STEMPEL mit tschechischen Input festzustellen, da nur im IR-Prozess selbst eine Verbesserung oder Verschlechterung durch den Einsatz des Stemmers zu ermitteln ist.

Eine Variante und mögliche Lösung für das Stemming-Problem der tschechischen Sprache könnten sprachunabhängige Stemmingverfahren, die auf statistischen Ansätzen beruhen, darstellen.

Die Analyse der Problembereiche und Besonderheiten des Tschechischen hat auch ergeben, dass das Verfahren der Kompositazerlegung im Tschechischen durch das geringe Auftreten der Komposita zu vernachlässigen ist.

Die in dieser Arbeit erzielten Ergebnisse stellen ein solides Grundgerüst an IR-Komponenten für weitere Arbeiten im Bereich WebCLEF und der tschechischen Sprache im IR. So wird bereits die im Rahmen dieser Arbeit erstellte Stoppwortliste in einer multilingualen Stoppwortliste bei WebCLEF zum Indexieren des EuroGOV-Korpus und der dazugehörigen Anfragen eingesetzt (vgl. Jensen 2005). Weiterhin schafft der intellektuell erstellte Text-Katalog für die tschechische Toplevel-Domain von WebCLEF einen Standard für die automatische Sprachidentifikation und ergänzt mit den Arbeiten von Jensen (2005) und Artemenko, Shramko (2005) die Basis für weitere Projekte im Bereich des Web-Retrievals.

7 Abkürzungsverzeichnis

Abb.	Abbildung
bspw.	beispielsweise
CLEF	Cross-Language Evaluation Forum
CLIR	Cross-Language Information Retrieval
ggf.	gegebenenfalls
Hrsg.	Herausgeber
IR	Information Retrieval
IRS	Information Retrieval-System
MIMOR	Mehrfachindexierung zur dynamischen Methoden-Objekt-Relationierung im IR
Tab.	Tabelle
TschNK	Tschechisches Nationalkorpus
u.a.	unter anderem
v.a.	vor allem
z.B.	zum Beispiel

8 Abbildungsverzeichnis

9 Tabellenverzeichnis

10 Inhalt der CD

* Magisterarbeit im PDF-Format

* Stoppwortlisten
    - Default-Stoppwortliste des Tschechischen Analyzer für Lucene
    - General Stoplist1, erstellt von Vašek Nemčík mit dem Korpus DESAM
    - General Stoplist2, erstellt von Vašek Nemčík mit den Korpora ESO und TschNK
    - Stoppwortliste des automatischen Sprachidentifizierers von Artemenko & Shramko
    - eigene Stoppwortliste

* STEMPEL
    - Der polnische Stemmer STEMPEL als ECLPISE-Projekt
    - Die Ergebnistabellen der 2 Durchläufe

* Der intellektuel erstellte Text-Katalog für die tschechische Toplevel-Domain von WebCLEF

* e-Mail-Auszug von Vašek Nemčík

11 Literaturverzeichnis

Verifizierungsdatum für die Interdokumente: 3. Juli 2005

Artemenko, O.; Shramko, M. (2005) *Entwicklung eines Werkzeugs zur Sprachidentifikation in mono- und multilingualen Texten*. Magisterarbeit Internationales Informationsmanagement, Universität Hildesheim.

Baeza-Yates, R.; Ribeiro-Neto, B. (Hrsg.) (1999) *Modern Information Retrieval*. Addison Wesley Longman Limited: Essex.

Belkin, N.J. (1984) *Cognitive models and information transfer*. In: Social Science Information Studies. 4, Pp. 111-129.

Belkin, N.J.; Croft, W. (1987) *Retrieval Techniques*. In : Williams, M. (Hrsg.) (1987) Annual Review of Information Science and Technology. Elsevier Science Publishers: New York, Pp. 109-145.

Bowker, L.; Pearson, J. (2002) *Working with Specialized Language. A practical guide to using corpora*. Routledge: London.

Bush, Vannevar (1945) *As we may Think*. The Atlantic Monthly.
http://ccat.sas.upenn.edu/~jod/texts/vannevar.bush.html

Bußmann, H. (1990) *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag: Stuttgart.

Carpineto et al (2001) *An Information-Theoretic Approach to Automatic Query Expansion*. In: ACM Transactions on Information Systems. Vol. 19. No. 1. January 2001, Pp. 1–27.

Český národní korpus - SYN2000. Ústav Českého národního korpusu FF UK, Praha 2000. http://ucnk.ff.cuni.cz

DUDEN – Das Herkunftswörterbuch (1997) bearbeitet von Drosdowski, Günther. Nach den Regeln der neuen deutschen Rechtschreibreform überarbeiteter Nachdruck der 2.Auflage. DUDEN Band 7. DUDENVERLAG: Mannheim.

Ferber, R. (2003) *Information Retrieval - Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*.dpunkt.verlag: Heidelberg.

Frakes, W.; Baeza-Yates, R. (1992) *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall : New Jersey.

Frisch, E.; Kluck, M. (1997) *Pretest zum Projekt German Indexing and Retrieval Testdatabase (GIRT) unter Anwendung der retrievalsysteme Messenger und freeWAISsf.* IZ-Arbeitsbericht 10. Informationszentrum Sozilawissenschaften: Bonn.

Fuhr, N.(1997): Skript zur Vorlesung *Information Retrieval*. Universität Dortmund. ftp://ls6-www.informatik.uni-dortmund.de/pub/doc/courses/ir/irmat.html

Galambos, L.(2004): *Semi-automatic stemmer evaluation*. IIPWM 2004.

Global Reach, Internet Statistics
http://global-reach.biz/globstats/index.php3

Grefenstette, G. (1998) *Cross-Language Information Retrieval*. Kluwer Academic Publisher: Massachusetts.

Hackl, René (2004) *Merhsprachiges Information Retrieval im Rahmen von CLEF 2003*. Magisterarbeit Internationales Informationsmanagement. Universität Hildesheim.

Haag, M. (2002) *Automatic Text Summarization. Evaluation des Copernic Summarizer und mögliche Einsatzfelder in der Fachinformation der DaimlerChrysler AG.* Shaker Verlag: Aachen.

Von Hahn, W. (2004) Skript zur Vorlesung *Computerphilologie- Themenfeld Korpusforschung.* Universität Hamburg. Wintersemester 2004/2005. http://nats-www.informatik.uni-hamburg.de/~vhahn/German/CP/Vorlesung0405/10aKorpus0405.pdf

Janich, N.; Greule, A. (Hrsg.) (2002) *Sprachkulturen in Europa - ein internationales Handbuch.* Gunter Narr : Tübingen.

Jensen, N. (2005): *Multilinguales Web Retrieval am Beispiel des EuroGOV Korpus*. Magisterarbeit Internationales Informationsmanagement, Universität Hildesheim.

Kluck, M.; Mandl, T.; Womser-Hacker, C. (2002) *Cross-Language Evaluation Forum (CLEF): Europäische Initiative zur Bewertung sprachübergreifender Retrievalverfahren.* In: Information – Wissenschaft und Praxis vol. 53 (2), Pp. 82-89.

Knorz, Gerhard (1995) *Information Retrieval-Anwendungen*. In: Zilahi-Szabo (Hrsg.) Kleines Lexikon der Informatik und Wirtschaftsinformatik, Pp.244-248.

Korfhage, R. (1997) *Information Storage and Retrieval*. Wiley: New York, 1997.

Kowalski, G. (1997) *Information Retrieval Systems: Theory and Implementation.* Kluwer
Academic Publishers: Boston/Dordrecht/London.

Kuhlen, R.; Griesbaum, J. (2001) Skript zur Vorlesung *Informationretrieval, Retrievalfunktion (Matching).* Department of Computer and Information Science at the University of Konstanz.
http://www.inf-wiss.uni-konstanz.de/CURR/winter0102/IR/v7_matching.pdf

Kuropka, D. (2004) *Modelle zur Repräsentation natürlichsprachlicher Dokumente. Ontologie-basiertes Informations-Filtering und –Retrieval mit relationalen Datenbanken.* In: Advances in Information Systems and Management Science. Band 10. Logos Verlag: Berlin.

Krcmar, Helmut (2000), *Informationsmanagement*, Springer Verlag: Berlin.

Lam-Adesina, A; Jones, G. (2001) *Applying Summarization Techniques for Term Selection in Relevance feedback.* In: *SIGIR'01.* September 9-12. 2001. New Orleans, Louisiana, USA.

Leopold, E. (2002) *Das Zipfsche Gesetz*. In: Joachims, T. & Leopold, E. (Hrsg.) *Schwerpunkt Textmining - Künstliche Intelligenz* 02/02. Fraunhofer Institut für Autonome intelligente Systeme: Sankt Augustin.
http://www.ais.fraunhofer.de/~leopold/Zipfkurz.pdf

Luckhardt, H.; Harms, I. (2005) *Virtuelles Handbuch Informationswissenschaft, Automatische und intellektuelle Indexierung.* Universität des Saarlans. http://is.uni-sb.de/studium/handbuch/index.php

Luhn, H. (1958) *The Automatic Creation of Literature Abstracts.* In: IBM Journal of Research and Development,Vol. 2 – No. 2.

McNamee, P.; Mayfield, J. (2002) *Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources*. SIGIR'02, August 11-15, 2002, Tampere, Finnland.

Mandl, T., Womser-Hacker, Ch. (2003) *Proper Names in the Multilingual CLEF Topic Set.* In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (Hrsg.) *Evaluation of Cross-Language Information Retrieval Systems.* Proceedings of the CLEF 2003 Workshop. Berlin et al. : Springer [Lecture Notes in Computer Science]. [to appear] Vorab in: Working Notes for the CLEF 2003 Workshop. 21-22-08.2003, Trondheim, Norway, pp. 439-443.

http://clef.iei.pi.cnr.it:2002/2003/WN_web/53.pdf

Mandl, T., Womser-Hacker, Ch. (2000) *Ein adaptives Information-Retrieval-Modell für Digitale Bibliotheken.* In: Knorz, G.; Kuhlen, R. (Hrsg.) *Informationskompetenz - Basiskompetenz in der Informationsgesellschaft.*Proceedings 7. Intl. Symposium für Informationswissenschaft. (ISI 2000). 8.-10.11.2000, Darmstadt.Konstanz: Universitätsverlag [Schriften zur Informationswissenschaft Bd. 38]. Pp. 1-16.

Mayfield, J. (2002) Präsentation zum Seminar *Introduction to Information Retrieval*. The Johns Hopkins University, Applied Physics Laboratory.
http://www.clsp.jhu.edu/ws2002/preworkshop/mayfield.pdf

Meadows, J. (1991) *Knowledge and communication*. Essays on the information chain. Library Association Publ. : London.

Meadows, C.; Boyce, B.; Kraft, D. (2000) *Text Information Retrieval Systems*. 2. Ausgabe. Academis Press, San Diego.

Robertson, S. (1981) *The methodology of information retrieval experiment.* In: Sparck Jones, K. (Hrsg.) *Information retrieval experiments*. Butterworths: London. Pp. 9-31.

Nohr, H. (2000) *Automatische Dokumentindexierung – Eine Basistechnologie für das Wissensmanagement.* In: Arbeitspapiere Wissensmanagement, Nr. 2/2000, Fachhochschule Stuttgart.
http://www.iuk.hdm-stuttgart.de/nohr/KM/KmAP/Indexing.pdf

Notes, G. (1999) *Search Engine Showdown*
http://www.notess.com/search/defs/

Oard, D. (1997) Cross-Language Information Retrieval Defined.
http://www.ee.umd.edu/medlab/mlir/mlir_definition.html

Paice, C. (1994) An *Evaluation Method for Stemming Algorithms*. In: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference. Springer Verlag: London, Pp. 42-50.

Peters, C. (Hrsg.) (2001) *Cross-Language Information Retrieval and Evaluation.* Workshop of the Cross-Language Avaluation Forum. CLEF 2000. Lisbon, Portugal, September 2000. Revised papers. Lecture Notes in Computer Science 2069. Springer: Berlin.

Poetsch, E. (2001) *Information Retrieval - Einführung in Grundlagen und Methode.* 2., völlig neu bearbeitete Auflage. Verlag für Berlin-Brandenburg: Potsdam.

Peters, C. (Hrsg.) (2001) *Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum CLEF 2000*, Lisbon, Portugal, September 2000. Revised Papers. Springer Verlag: Berlin, Heidelberg.

PRAVICA Sprachendienst
http://www.pravica.hr/ger/sprachen.html

Qiu., Y. (1995)*Automatic Query Expansion Based on A Similarity Thesaurus.* PhD Thesis, Swiss FederalInstitute of Technology (ETH). In: The Eurospider Retrieval System and the TREC-8 Cross-Language Track Martin Braschler, Min-Yen Kan, Peter Schäuble, Judith L. Klavans Eurospider Information Technology AG ZürichSwitzerland
http://trec.nist.gov/pubs/trec8/papers/eit_t8f.pdf

Qiu, Y.; Frei, H. (1993) *Concept Based Query Expansion*. In: Korfhage, R. (Hrsg.), Proc. o. t. Conf. on R & D in IR, Band 16 von ACM (SIGIR). Springer: Pittsburgh, S. 160-169.

Van Rijsbergen (1979): *Information Retrieval*. London(UK): Butterworths.

Salton, G.; McGill, M. (1987) Information Retrieval-Grundlegendes für Informationswissenschaftler. McGraw-Hill: Hamburg.

Salton, G.; McGill, M. (1983): *Introduction to Modern Information Retrieval*.McGraw-Hill: New York.

Sojka, P; Kopeček, I.; Pala, K. (Hrsg.) (2004) *Text, Speech and Dialogue.* 7th International Conference, TSD 2004 - Brno, Czech Republic, September 2004 - Proceedings. Springer-Verlag: Berlin.

Sullivan, D. (1999). *Search Engine Watch*.
http://www.searchenginewatch.com/facts/glossary.html)

Tague-Sutcliffe, J. (1995) *Measuring Information - An Information Services Perspective.* Academic Press : San Diego.

TREC-4 Proceedings.
http://www–nlpir.nist.gov/TREC/t4_proceedings.html

WebCLEF (2005) *Participation Guidelines*, –FINAL–, May 15, 2005, http://ilps.science.uva.nl/webclef/participants-guidelines-final-20050515.pdf

Will, C. (1993) *Comparing Human and Machine Performance for Natural Language Information Extraction: Results for English Microelectronics from the MUC-5 Evaluation*. In: Proc. of the Fifth Message Understanding Conference. Kaufmann Publishers: Morgan, pp. 53-67.

Witten, I.H., Moffat, A., Bell, T.C. (1994). *Managing Gigabytes: Compressing and Indexing Documents and Pictures*. Morgan Kaufmann Publishing, San Francisco.

Schwarz, C.; Thurmair, G. (Hrsg.) (1986) *Informationslinguistische Texterschließung*. Georg Olms Verlag Hildesheim: Hildesheim.

Šlosar, D. (1999) *Česká Kompozita Diachronně,* Masaryková Univerzita: Brno.
(Auflistung der Komposita S. 106-124)

Womser-Hacker, C. (1996) *Das MIMOR-Modell. Merfachindexierung zur dynamischen Methoden-Objekt-Relationierung im Information Retrieval.* Habilitationsschrift. Universität Regensburg, Informationswissenschaft.

Womser-Hacker, C. (2003) Skript zur Vorlesung *Einführung in die Informationswissenschaft*. Universität Hildesheim, Wintersemester 2003/2004, http://www.uni-hildesheim.de/~womser/Lehre/IW_Einfuehrung.htm